nature

www.nature.com/nature

Vol 445 | Issue no. 7126 | 25 January 2007

Grounds for optimism

A summit of Africa's leaders marks a deepening commitment to science and technology in the continent.

Next week, heads of states in the African Union will gather for a summit in Addis Ababa (see page 356). The meeting's two main themes are climate change and the harnessing of science and technology for development — making this Africa's highest-powered gathering on science in years.

Issues to be discussed include a proposal for an Africa-wide science fund (broadly analogous to the European Union's Framework programme), plans for new research centres, and a biotechnology strategy for Africa. This last issue is an attempt to reconcile supporters and critics of agricultural biotechnology and establish some level of consensus on research, commercialization and regulation. Also on the table is an idea to establish a presidential science council, at which Africa's heads of state would be briefed on relevant issues in science and development, and review African Union decisions on science policy.

These proposals will be vigorously debated at the Addis summit. Some countries — notably Nigeria and South Africa, the two nations with the most money — are still to be convinced of the need for an Africa-wide science fund. At the same time, civil servants and science ministers across the continent are nervous about the idea of a presidential council, as they perceive (not unreasonably) that it could interfere with their own organization, the newly established African Ministerial Council on Science and Technology, also a component of the African Union.

The idea for a presidential council is not new: it was also put forward and agreed at the last African leaders' summit to deal extensively with science almost two decades ago. The inspiration that time came from Kenyan entomologist Thomas Odhiambo, who used his legendary charm and influence to persuade heads of state that they should meet periodically with scientists. One of his strongest supporters was Nigeria's current elected president Olusegun Obasanjo, who was then the nation's military ruler. Nigeria has the largest population of any nation in Africa — around 130 million — and its oil revenues have increased its potential to exert influence across the continent.

The rationale behind the presidential council is straightforward: if you want to get anything done, you need access to the people at the top. Yet the omnipotence of typical participants in the 1980s proved to be the idea's undoing last time, after just three meetings between scientists and presidents. "You couldn't tell who was about to be overthrown in a coup, or who was next in line to be killed. There was no continuity in our work," one of the organizers tells *Nature*.

Times have changed. Obasanjo's keenness on science and innovation is shared by several of his fellow elected presidents, notably Abdoulaye Wade of Senegal, Bingu wa Mutharika of Malawi, and Rwanda's Paul Kagame. Many more countries in Africa also have elected governments, resulting in the healthy involvement of fresh constituencies in scientific decision-making.

Nowhere is this more apparent than in the build-up to next week's summit. Representatives of many different interests, including scientists, non-governmental organizations, civil servants, agencies of the United Nations, philanthropies and the media, have been free to scrutinize and influence the summit's agenda. They were encouraged to do so by the African Union, which is loosely based on the European Union and may be much better placed than its predecessor — the Organization of African Unity — to push forward meaningful initiatives in science and other spheres.

"The planners of the summit want Africa's leaders to think about creating conditions in which war and poverty are the exception, rather than the rule."

But international politics could still throw the summit's plans to attend to scientific matters off course. Ethiopia, the summit's host nation, is currently engaged in a contentious military intervention in neighbouring Somalia, and the leaders' response to that situation could yet come to dominate the meeting.

So far, however, the organizers of the Addis summit have conceded no changes in its agenda, reflecting their determination that although the Somalia situation must be discussed, it should not dominate proceedings. This is as it should be. The planners of the Addis summit want Africa's leaders to think about the long term, and about creating conditions within which war and poverty are the exception, rather than the rule. They should not be distracted from that goal. ■

Clock-watching

Time for a change?

The clock face around which our minutes tick away is perhaps the single most potent symbol of the industrial era. It is the clock that allowed modern life to be cut precisely into the segments that the workplace requires and the individual seeks to protect.

On wrists and office walls, hanging from the vaults of railway stations, or squatting toadlike by the bedside to curtail our sleep, the clock and its face have become the near-universal embodiment of the always felt but hitherto seldom quantified march of time. As a result, it cries out for manipulation — whether by the hurried commuter who fools himself into punctuality by setting his watch five minutes fast, or by a government changing the clocks of a nation to make the most of the daylight hours.

There was a time — 1911, as it happens — when this journal was strongly against the latter practice, as embodied in the then radical idea of introducing daylight-saving time to Britain. Rather high-mindedly, we thought that "the scheme is unworthy of the dignity of a great nation, and if it were made compulsory by legislation, it

would be a monument to national flaccidity". In the subsequent 95 years, however, this stance has softened, and *Nature* welcomes the proposal currently before the British parliament to extend current daylight saving by putting Britain's clocks forward by one hour all year round. This would put Britain into the GMT+1 time zone in winter and GMT+2 in summer, bringing the nation into line with the rest of Europe.

The evidence is that lighter evenings make life safer and may well save energy too (see page 344). A three-year experiment along these lines, well monitored to ensure that the change lives up to its proponents' claims, seems a sensible idea.

Much less persuasive is a separate plan to move forward the time on the 'Doomsday clock', an icon cannily created by the editorial board of the *Bulletin of the Atomic Scientists* in 1947 to alert the world to the looming threat of nuclear annihilation. Last week the current board decided to broaden the clock's remit to include such developments as climate change, and to move its hands forward by two minutes to just five minutes to midnight.

This raises two concerns. A minor one is that 'nuclear war' means something rather different today to what it meant at the time of, say, the Cuban missile crisis (when the hands were set further from midnight than they are now). A nuclear war is no longer necessarily synonymous with an all-out exchange between superpowers, and may not lead to the doomsday envisaged by the creators of the doomsday clock. This is not to minimize the horrors of a limited nuclear exchange, but to acknowledge that the context for the still vital project of averting any military use of nuclear weapons has changed.

A greater misgiving comes from the addition of non-nuclear concerns to the doomsday calculus. This seriously muddies the waters. Climate change is undoubtedly a major challenge, but it does not threaten doomsday in the manner of a full-blown nuclear war. Global warming has no hair trigger, no tiny margin between safety and disaster, no doom that can be unleashed in the flight time of a missile — none of the characteristics, in fact, that made the fatal minutes on the face of the doomsday clock so iconic.

Climate change is a substantial threat, but it is quite different in character to nuclear war: it is the deterioration of land, the increase of drought, a billion livelihoods descending from backbreaking to impossible. The principal human cost of climate change is likely to be an intensification of global mortality due to poverty and ill health — mortality that already runs at a level that all would condemn as unacceptable were it not that, as a world, we accept it. This moral weight makes it pressing, but does not make it urgent in the 'time ticking away' sense the doomsday clock so powerfully evokes. It is more important that policies to reduce the harm done by climate change be sustainable over the long term than that action be taken precipitously.

To fight climate change, we do not need to alarm ourselves with clocks of doom. Instead we just need to use our time to good purpose. And the reduction in energy use to be expected from single–double daylight saving in Britain — or from the extended single daylight saving that is to be implemented in the United States this year — will be a marginal, but nonetheless welcome, step in the right direction. ■

Making connections

A series of essays is launched in *Nature*.

There are times in the development of science when a shift in approach is sufficiently extensive that only a collection of thoughts and perspectives from many different practitioners can do justice to it. This issue sees the launch of a themed series of essays, called Connections, which take such an approach to the interdisciplinary study of complex, dynamic systems (see page 369).

Scientists in almost every discipline are grappling with the problem of how best to model such systems. Cell biologists are being driven to do so, for example, by the surge of data from techniques that reveal biological processes in unprecedented detail, and quantum physicists by properties exhibited by collections of particles that would not have been anticipated on the basis of how a single particle behaves. Across these fields and beyond, deeper insight requires a systems-level approach that seeks to understand interactions and make connections. Although the goal is clear enough, the way to reach it is not.

Many researchers recoil from terms such as 'systems biology' and 'complexity,' interpreting them as euphemisms for things we don't adequately understand. Research on problems involving dynamic interactions between large numbers of entities is often directed by the availability of the data, rather than by a carefully considered question. And a rush of studies claiming to uncover simplifying principles that unite complex networks has sometimes generated more heat than light.

In some cases, attempts to understand networks and whole systems are driving researchers to cross disciplinary boundaries. Social and physical scientists are often more accustomed to such collective activity than molecular biologists, for whom the borrowing of techniques and expertise is now becoming commonplace.

The essays in the Connections series will illustrate some of the insights that are emerging as researchers pursue more holistic approaches to problems, while engaging in an unprecedented degree of collaboration between biological, social and physical scientists. They will raise provocative ideas about how to probe dynamic systems, illustrating, for example, how systems approaches can challenge assumptions established within the more reductionist framework of twentieth-century science.

"Deeper insight requires a systems-level approach that seeks to understand interactions and make connections."

The series begins just a few weeks after the popular Essay page returned to *Nature*. It will reiterate the tradition of this format as a forum for scientists to reflect on new ideas, or re-evaluate old ones. The Connections essays will also be collated on the web, where access to the first four will be free. We hope that, week by week, a greater story develops than could be told by a single article — and that over the coming year, the Essay page will continue to provide scientists with a valuable opportunity to say exactly what they think. ■

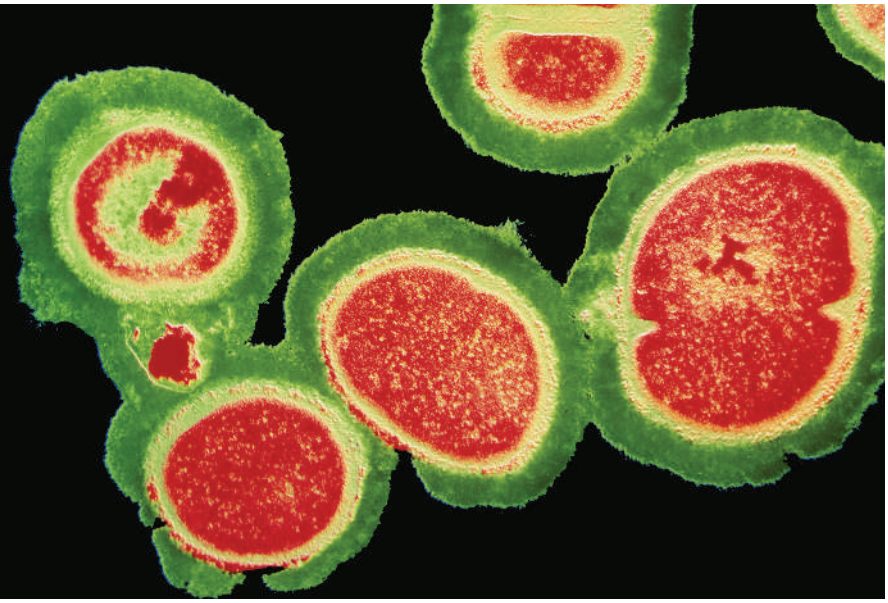
RESEARCH HIGHLIGHTS

MRSA toxin characterized

Science doi:10.1126/science.1137165 (2007)
Researchers have characterized a toxin secreted by some strains of the menacing bacterium methicillin-resistant *Staphylococcus aureus* (MRSA, pictured).

Antibiotic-resistant *S. aureus* has plagued hospitals for decades and is now showing up more frequently outside hospitals too.

A team led by Gabriela Bowden of Texas A&M University in Houston, and François Vandenesch of the University of Lyon, France, studied the toxin 'Panton Valentine leukocidin', which is produced by some MRSA strains that cause a severe form of pneumonia. The toxin alone could cause pneumonia in mice. The team found that it also stimulates expression of two bacterial proteins: one that promotes inflammation, and another that may increase the bacterium's ability to stick to injured airways in the lungs.



K. LOUNATMAA/SPL

OPTICS

Stability is everything

Phys. Rev. A **75,** 011801 (2007)
Even the tiniest vibration can upset the time-keeping of an atomic clock, so physicists endeavour to make its parts stable against acceleration. Stephen Webster of the UK National Physical Laboratory in Teddington and his colleagues report progress for one component: the 'optical cavity'.

The optical cavity helps to fix the frequency of the laser light that interrogates the atoms of the clock. It consists of a cylinder with a mirror positioned at each end. The researchers analysed the mechanics of this set-up, finding that they could minimize a cavity's sensitivity to vibrations by slicing some material from the cylinder's lower side and by carefully positioning the cylinder's supports. This helped them to build a cavity 15 times less sensitive to vertical accelerations than the best existing design.

MOLECULAR BIOLOGY

Through the tunnel

Nature Struct. Mol. Biol. doi:10.1038/nsmb1197 (2007)
Cholesteryl ester transfer protein (CETP) works to lower levels of 'good' cholesterol and raise levels of 'bad' cholesterol; in theory, blocking its action should help boost the good stuff. But in December 2006, Pfizer pulled the plug on trials of a drug (torcetrapib) designed to do this, because of concerns about patient safety.

Xiayang Qiu and his colleagues, working at Pfizer in Groton, Connecticut, have now determined the crystal structure of CETP. The protein contains a tunnel through which, they propose, cholesterol and other lipids are shunted. They want to understand how torcetrapib blocks this process. This may help researchers to find other CETP inhibitors that avoid torcetrapib's problems.

NUTRIENT CYCLING

It's simple, really

Science **315,** 361–364 (2007)
The complex world of biogeochemistry does at least contain some simplicity. A ten-year study has shown that the rate at which nitrogen is released from decomposing plant matter depends on just a couple of variables, regardless of climate or ecosystem type.

J. SEXTON

A team led by William Parton of Colorado State University in Fort Collins placed leaves (pictured below) from seven different ecosystem types at 21 sites worldwide, and left them to decompose. They found that the rate at which nitrogen compounds are released to the soil depends almost entirely on the initial concentration of nitrogen in the plant matter, and the leaf and root mass remaining at a given time. Only in arid grasslands did other factors come into play.

QUANTUM COMPUTATION

Slow down a bit

Phys. Rev. Lett. **98,** 020501 (2007)
Will quantum computers work if they take a long time to read the data?

In modelling the operation of quantum computers, physicists have tended to assume that a quantum bit's state can be read as quickly as the value of the state could be changed. But it seems likely that measurements in real devices will be comparatively slow. Might this derail the error-correction procedures needed to keep a quantum computer running by allowing noise to swamp the data?

Happily, David DiVincenzo of IBM's T. J. Watson Research Center in Yorktown Heights, New York, and Panos Aliferis of the California Institute of Technology in Pasadena have devised a new error-correcting protocol that can tolerate slow measurements. With this protocol, even measurements that take 1,000-fold longer than state manipulations will have little effect on the computer's accuracy.

S. PHILLIPS, HARVARD UNIV.

BIOTECHNOLOGY

Looks good on paper

Angew. Chem. Int. Edn doi:10.1002/anie.200603817 (2007)

Paper printed with polymer channels could form a cheap and portable testing lab for biological samples, say scientists led by George Whitesides at Harvard University in Cambridge, Massachusetts.

To make the simple bioassay device, the researchers first soaked chromatography paper in a solution that forms a water-repellent polymer when exposed to ultraviolet light. They then used a mask to expose only certain regions to the ultraviolet light, defining channels and test areas. In the demonstration device (pictured right), the test areas were primed with colour-changing reagents to detect glucose (left) and protein (right). Liquid is drawn through the channels by capillary action when the edge of the paper is dipped into the sample.

NANOTECHNOLOGY

Reinventing the wheel

Nature Nanotech. doi:10.1038/nnano.2006.210 (2007)

Efforts to build machinery on the nanoscale are rolling forward, with new work reporting a molecular wheel.

Leonhard Grill of the Free University of Berlin, Germany, and his colleagues show that triptycene groups, which resemble three-bladed paddlewheels, can act as wheels only 8 angstroms wide. They fixed one triptycene to each end of a rigid axle, then pushed this primitive molecular vehicle over a copper surface using the tip of a scanning tunnelling microscope.

Evidence that the wheels could 'roll' came from looking closely at how the vehicle

moved. When one of its three-armed wheels rotated by 120°, the molecule jumped and its axle reoriented. Nanowheel rotation has been claimed before, but never shown directly.

GENETICS

Hand-me-down cells

Proc. Natl Acad. Sci. USA doi:10.1073/pnas.0606169104 (2007)

The idea that a mother would give anything to protect her child has been extended by new research.

J. Lee Nelson of the Fred Hutchinson Cancer Research Center in Seattle, Washington, and her colleagues found that children with type 1 diabetes have higher levels of their mother's DNA in their blood than do their unaffected siblings, implying that they inherit more maternal cells. They also found small populations of female insulin-producing cells in male pancreases.

Putting the two observations together suggests that a mother's insulin-producing cells, transferred to the fetus in the womb, could contribute to the regeneration of her child's damaged pancreatic cells.

GEOLOGY

Fast movers

Geology **35,** 29–32 (2007)

Plate tectonics was more changeable in the past than once believed, a new study suggests.

Earlier work indicated that the rate of new crust being born at seafloor spreading ridges has stayed relatively constant over the past 100 million years or so. To check this, Clinton Conrad of the Johns Hopkins University in Baltimore, Maryland, and Carolina Lithgow-Bertelloni of the University of Michigan in Ann Arbor analysed afresh the ages of sea floors in different ocean basins.

They find that, globally, rates of seafloor spreading increased by about 20% between 60 million and 30 million years ago. Since then, because a fast-spreading system in the Pacific has been recycled into Earth's depths, the average spreading rate has dropped by 12%.

CELL BIOLOGY

A protective pair?

Neuron **53,** 233–247 (2007)

A molecular link between two signalling pathways in the central nervous system has been uncovered by Stephen Moss of the University of Pennsylvania in Philadelphia and his co-workers.

The team shows that an enzyme known as AMPK, implicated in appetite signalling, interacts with the neuronal receptor $GABA_B$. AMPK seems to add a phosphate group to the receptor, modifying its activity.

The finding may help to pin down AMPK's role in the brain's response to injury. The enzyme is activated in stressed tissue, but there is conflicting evidence about whether it prevents or exacerbates neuronal damage. Moss's group suggests that $GABA_B$ mediates a neuroprotective effect.

JOURNAL CLUB

Daniel Pauly
University of British Columbia, Vancouver, Canada

A marine biologist dives into the history of the Gulf of California.

A decade ago, I coined the term 'shifting baselines' to describe how society perceives environmental change. The concept has caught on: there's even a website, at www.shiftingbaselines.com, featuring short, explanatory films.

The films push the idea that the standards by which society assesses change are themselves changing. We tend to use the state of affairs that prevailed when we first became aware of an issue as our reference point for evaluating future change — a baseline that shifts with each generation.

A set of three brilliant papers illustrates how this can shape our understanding of ecosystems.

The most recent paper (A. Sáenz-Arroyo *et al. Fish Fish.* **7,** 128–146; 2006) reconstructs from historical sources, such as pirates' logs, details of the Gulf of California's ecosystem stretching back to the sixteenth century. The researchers argue that the past abundance of creatures such as marine mammals, turtles and oysters recounted in these sources should be considered when setting conservation targets today.

Their previous work examined records of Gulf groupers, fish that once dominated the area's reefs (A. Sáenz-Arroyo *et al. Fish Fish.* **6,** 121–133; 2005), concluding that fishery statistics didn't go back far enough to accurately map the species' decline.

Further, they quizzed three generations of artisanal fishers (A. Sáenz-Arroyo *et al. Proc. R. Soc. Lond. B* **272,** 1957-1962; 2005), and found that fishers' knowledge of the location or habits of species disappeared within one generation, if the species became rare.

We are all affected by this kind of collective amnesia. It allows us to handle change. But it is also the reason why we accept losses that would be intolerable, were we aware of them.

SPECIAL REPORT

Saving time

Politicians in the United Kingdom and United States have launched efforts to extend daylight-saving measures — hoping to save lives, cut power use and combat carbon emissions. But energy experts say that it's not that easy. **Michael Hopkin** reports.

The UK bill to bring the clocks forward could change the face of London's time.

A. WOOLFITT/ROBERT HARDING WORLD IMAGERY/CORBIS

On 26 January, the British parliament will vote on whether to turn its back on a cherished piece of the country's heritage: Greenwich Mean Time, long the standard against which the world's clocks were set. Under the proposal, Britons would set their watches forward an hour year-round, putting them on the same time as the rest of Europe, and never on Greenwich time.

If adopted, the measure would give Britain an extra hour of daylight in the evening. Backers argue that the proposal could cut automobile accidents, saving more than 100 lives per year, and slash demands on the country's power grid, saving carbon emissions equivalent to those generated by 70,000 people in a year.

British clocks would still go forward an hour every spring and back again in autumn, just as they have almost every year since 1916, when the country adopted British Summer Time. Under the new scheme, which would be trialled for a period of three years, clocks would be set an hour ahead of Greenwich in the winter, and two hours ahead in summer — a regime called 'single-double summer time'.

But history shows that such seemingly simple changes are not made easily. A similar experiment in the late 1960s failed, mainly because of parents' complaints about children going to school in the dark and the grumblings of farmers and other outdoor workers.

Energy experts say that it's not so easy to calculate the true benefits for the environment. The move comes as US citizens prepare for a longer summer, by pushing their clocks forward on 11 March — three weeks earlier than usual. US politicians have claimed that the country will save 100,000 barrels of oil per day, but will wait until the scheme is underway to gather detailed data on changes in energy use.

Both the British and US policies are extensions of the current scope of daylight-saving time (DST), which is practised in roughly 70 countries, almost all of which are at temperate latitudes. DST artificially shifts the time of dawn and dusk forward by an hour during the long summer days, providing an extra hour of light in the evenings. As the winter days draw near, the clocks are put back again so that mornings don't become unbearably dark.

But the DST scheme can be taken further, argues Elizabeth Garnsey, an innovation researcher at the University of Cambridge, UK, who has evaluated the probable effects of the proposed changes to the British system. "Countries are starting to realize that their daylight-saving policies haven't really been saving daylight," she says.

Moving the clocks forward yet another hour would produce a slew of benefits, she says. The reduced need for lighting in the afternoons could save around £485 million (US$957 million) a year, as well as 170,000 tonnes of carbon dioxide. Preliminary calculations cited by the proposal's backers suggest that domestic lighting bills could dip by 0.8%, whereas commercial spending could be cut by 4% as more working days finish during daylight. And because the late afternoon is also the time of peak power demand, analysts at National Grid, which coordinates Britain's power supply,

CLIMATE SATELLITES
Researchers detail their wish list to avoid 'fatal' gaps in measurement.
www.nature.com/news

Turning back time

Daylight-saving time (DST) has a long and chequered history — popular in some time periods, not so in others.

1784 — The idea of shifting clocks to make better use of daylight hours is first proposed by Benjamin Franklin in a satirical letter to the *Journal of Paris*. He later whimsically suggests that people be taxed for having their shades drawn and candles lit during daylight hours.

1907 — The scheme is proposed in earnest by UK builder and businessman William Willett, in a pamphlet entitled *The Waste of Daylight*.

1916 — Germany becomes the first country to adopt DST, partly to conserve resources for its war effort. Britain follows suit three weeks later. The editors of *Nature* scornfully suggest that one might also alter thermometers to register 'warmer' temperatures in winter.

1917 — Newfoundland becomes the first North American region to adopt DST.

BETTMANN/CORBIS

1918 — The United States adopts DST (see left) and divides the country into time zones. This change is made despite the protestations of Congressman Otis Wingo that "while our boys are fighting in the trenches, we are here like a lot of schoolboys tinkering with the clocks". DST is dropped the following year.

1948 — Japan experiments with DST at the behest of the occupying Allied army. It drops the practice in 1951 and to this day is the only major developed nation without DST.

1966 — The United States formally reinstates DST after experimenting with it during the Second World War. The partial success of 'war time' had led to a disorganized system in which states and even individual counties were free to decide how and when to implement DST.

1968 — Britain begins its ill-fated attempt to abandon twice-yearly clock changes and observe 'summer time' throughout the year. The scheme collapses in 1971 under pressure from disgruntled early risers in the construction and farming industries.

predict that fewer power plants will be needed at any one time, so less efficient ones will be needed less often.

The problem is, however, that the science behind the numbers remains sketchy at best. For her calculations, Garnsey and her colleagues compared the energy consumption only in the weeks directly on either side of a clock change, as the gradual shifting of the day length over the year makes it very difficult to compare annual energy use on regular time versus DST. And so it remains hard to say what the overall year-round effect would be. "I don't think that any really exhaustive work has been done anywhere," Garnsey admits. "In terms of carbon emissions, I wouldn't say we really know yet. But in spring and autumn, there is quite substantial waste of electricity."

"In spring and autumn, there is quite substantial waste of electricity."

Another main argument for pushing the clocks forward has been the estimated lives saved from automobile accidents. Lord Sainsbury, the former British science minister, has claimed that the earlier experiment, when the country remained an hour ahead of Greenwich time year-round between1968 and 1971, saved an average of 140 lives a year. That adds up to nearly 5,000 needless deaths over the past quarter-century as a result of adhering to Greenwich time in winter, says Garnsey. "That should be the deciding factor," she says.

Others point to the increased opportunity for the old and the young to pursue outdoor activities in the evening, and benefits to businesses of aligning themselves with continental Europe. "The quality of life of nearly everyone will be improved," says Tim Yeo, the member of parliament who is sponsoring the UK bill.

Previous experiments with daylight-saving changes have not exactly borne that out. In Australia in 2000, officials experimented with an early start to DST to save energy. But energy consumption in Sydney, the host city for the Olympic Games that year, did not really fall and the change was not made permanent.

And two of the largest developing countries in terms of energy — India and China — do not observe DST, even though they cover huge areas; China, in particular, spans four time zones yet has the same time across the country and all year round.

In the United States, Congress approved the shift in their DST last year. The clocks will be put forward in March rather than April, then back in November rather than October. The change gains Americans four extra weeks of DST, the first change in the system since 1966. Most of Canada will observe the same change.

Meanwhile, the outcome of the British legislation hangs in the balance. Yeo, the measure's most prominent supporter, is a member of the opposition Conservative party, and the Labour government has refused to back the idea. The issue is more about politics for popularity than sound policy, argues energy expert Mayer Hillman, emeritus fellow at the Policy Studies Institute in London. He has been calling for Britain to adopt single-double summer time since 1988. Although lighter evenings led to a drop in the number of road accidents, he notes that a popular myth persists that putting the clock forward is more dangerous, because a small spike was seen in the number of early-morning accidents after DST was introduced. "I can't tell you how maddening it is," he says.

And some might doubt that the precise hours of daylight really make a difference to people's behaviour anyway. A new survey of more than 21,000 Germans (T. Roenneberg, C. J. Kumar & M. Merrow *Curr. Biol.* **17,** R44–R45; 2007) suggests that, in rural areas at least, people naturally fall into synch with the different daylight hours as they vary with geographical location, rising and going to bed with the Sun irrespective of the time shown on their watch.

If the US and British proposals prove successful, then analysts can finally come closer to answering questions about the true importance of setting the clocks correctly. "It may be that in the middle of winter there's not really much benefit," says Garnsey. "But the reason the experiment is needed is to get the correct data. There hasn't been an extensive analysis — how much energy will you save?" ■

See editorial, page 339.

How to drive light round the wrong bend

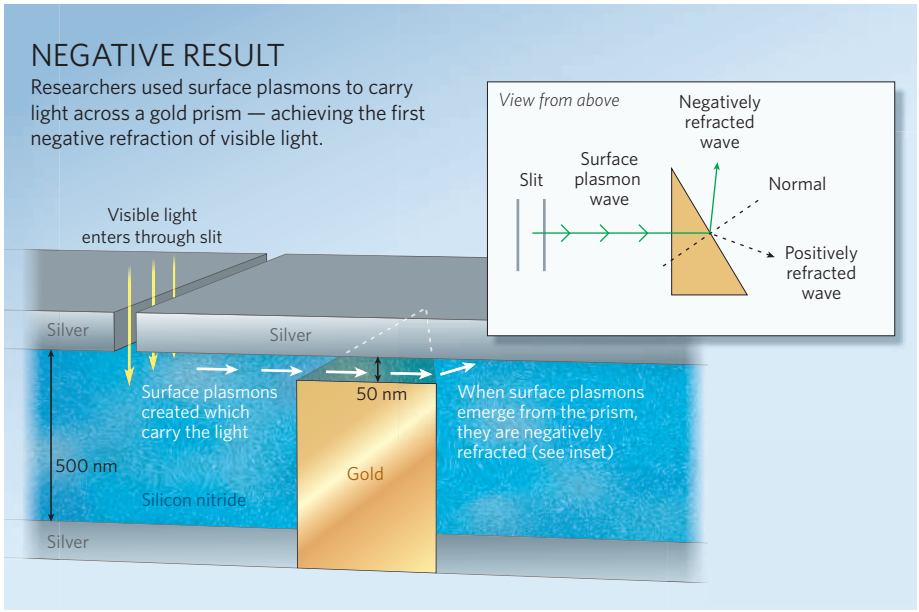
Can visible light ever be manipulated so that it bends the wrong way? If it could, a range of futuristic devices would be tantalizingly close to reality, such as a lens for imaging features smaller than the wavelength of light, or a shield to render objects invisible.

Several scientists have written off such 'negative refraction' in the visible range as practically impossible but a group is now claiming to have achieved it, spurring a debate about what constitutes true refraction.

Light bends in a specific way when it passes from one medium to another — an effect called refraction. Negative refraction describes a situation in which light bends the opposite way. It happens only if the direction in which the peaks and troughs travel along a light wave can be reversed relative to the direction in which the wave itself is travelling.

A material with a negative refractive index would focus light perfectly instead of dispersing it. This led John Pendry of Imperial College London to predict that a 'perfect' lens could be made, which would image features smaller than the wavelength of light. Some asserted that refraction could only ever have a positive value. But the debate was settled in 2003 when negative refraction was demonstrated for microwaves[1,2] and later for infrared waves.

Researchers achieved the effect with 'metamaterials' that had components of roughly the same size as the light's wavelength. More recently, Pendry used a metamaterial to bend light around an object to create an 'invisibility shield', also for microwaves[3].



But achieving similar effects for visible light has seemed well out of reach. Radiation in the microwave and infrared ranges has wavelengths in the order of micrometres or centimetres, so the components of the material used to negatively refract them are also on this scale. But building something equivalent for visible light, with a wavelength of some 500 nanometres, is a huge challenge.

Now Jennifer Dionne and Henri Lezec, working in Harry Atwater's group at the California Institute of Technology in Pasadena, have unveiled a material that they say has a negative refractive index for visible light. Dionne presented the results on 11 January at Nanometa 2007, a conference on nanophotonics and metamaterials held in Seefeld, Austria, and the group has submitted them for publication.

Rather than try to create a material with components as small as the wavelength of visible light, theoreticians recently suggested taking advantage of electromagnetic waves called surface plasmons, created when light hits free electrons oscillating on the surface of a metal, to guide the light in the desired direction. This is what Dionne and Lezec have now done. Their device, called a waveguide, consists of the insulator silicon nitride sandwiched between two sheets of silver.

Light enters the device through a slit in the upper silver sheet. Once inside, the light wave couples with oscillating electrons in the silver to create a surface plasmon wave that travels along the metal's surface. But embedded in the silicon nitride is a gold-coated prism, with a gap between it and the upper silver sheet that is just 50 nanometres wide (see graphic). As the surface plasmon wave crosses this gap, it is refracted. Dionne says that she has detected light with wavelengths of 480–530 nm (blue-green) emerging from the device having undergone negative refraction. The refractive index reached as low as −5 (compared with +1.33, for light travelling from air into water).

Passing the surface plasmons through the thin gap above the prism confines their move-

E. SCHREMPP/SPL

A straw in a glass of water seems disjointed because of refraction (left). But in this rough mock-up of what would happen if water had a negative refractive index (right), the effect is startling. The underside of the water's surface can be seen but not the bottom of the glass. For more accurate models, see ref. 4.

ment, so only one mode of surface plasmon wave can get through. At certain wavelengths of light, the frequency of the surface plasmon wave is close to the frequency of the oscillating electrons within the bulk of the metal. In this case the surface plasmon wave and the oscillating electrons interact in such a way that the direction of travel of the wave's peaks and troughs is reversed, giving negative refraction.

For Dionne, the goal of "peeking round the corner" has been achieved. "It's like alchemy," she says. "But it works."

Others in the field are more cautious. Mark Stockman, a theoretician at Georgia State University in Atlanta, is concerned about the system's inefficiency, pointing out that only about 1% of the light gets through. Dionne emphasizes that enough light gets through to be detected directly and says she thinks improvements can be made.

And some are unconvinced that it offers true negative refraction. Allan Boardman, a theoretician from the University of Salford, UK, and Vladimir Shalaev from Purdue University in West Lafayette, Indiana, who are also trying to negatively refract visible light, argue that the experiment simply shows negative refraction of plasmons, rather than of light itself. "It's not negative refraction per se," says Boardman. "They've got to qualify it a lot more."

But others such as Nikolay Zheludev of the University of Southampton, UK, say this doesn't really matter, because the end result is the same. "If everything is correct, this is a grand claim," says Zheludev. "Yes, they had negative refraction," agrees metamaterials and plasmonics expert Eli Yablonovitch from the University of California, Los Angeles. "I don't see much controversy there."

Pendry is also convinced, although he says he didn't expect to see the effect demonstrated so soon. "It is very impressive," he says. "They've done it in a most spectacular way."

Whether the approach counts as true negative refraction or not, to do anything useful with it will require turning the two-dimensional system into a three-dimensional device. Atwater envisages stacking a dense array of waveguides on end: "We have not done this yet, but at least this work illustrates the inherent possibility of doing so." ■

Katharine Sanderson

1. Parazzoli, C. G., Greegor, R. B., Li, K., Koltenbah, B. E. C. & Tanielian, M. *Phys. Rev. Lett.* **90,** 107401 (2003).
2. Houck, A. A., Brock, J. B. & Chuang, I. L. *Phys. Rev. Lett.* **90,** 137401 (2003).
3. Schurig, D. *et al. Science* **314,** 977–980 (2006).
4. Dolling, G., Wegener, M., Linden, S. & Hormann, C. *Optics Express* **14,** 1842–1849 (2006).

D. MOWBRAY, CIMMYT

WHEAT FUNGUS SPREADS OUT OF AFRICA
Stem rust threatens key crops in Asia.
www.nature.com/news

PR's 'pit bull' takes on open access

The author of *Nail 'Em! Confronting High-Profile Attacks on Celebrities and Businesses* is not the kind of figure normally associated with the relatively sedate world of scientific publishing. Besides writing the odd novel, Eric Dezenhall has made a name for himself helping companies and celebrities protect their reputations, working for example with Jeffrey Skilling, the former Enron chief now serving a 24-year jail term for fraud.

Although Dezenhall declines to comment on Skilling and his other clients, his firm, Dezenhall Resources, was also reported by *Business Week* to have used money from oil giant ExxonMobil to criticize the environmental group Greenpeace. "He's the pit bull of public relations," says Kevin McCauley, an editor at the magazine *O'Dwyer's PR Report*.

Now, *Nature* has learned, a group of big scientific publishers has hired the pit bull to take on the free-information movement, which campaigns for scientific results to be made freely available. Some traditional journals, which depend on subscription charges, say that open-access journals and public databases of scientific papers such as the National Institutes of Health's (NIH's) PubMed Central, threaten their livelihoods.

From e-mails passed to *Nature*, it seems Dezenhall spoke to employees from Elsevier, Wiley and the American Chemical Society at a meeting arranged last July by the Association of American Publishers (AAP). A follow-up message in which Dezenhall suggests a strategy for the publishers provides some insight into the approach they are considering taking.

The consultant advised them to focus on simple messages, such as "Public access equals government censorship". He hinted that the publishers should attempt to equate traditional publishing models with peer review, and "paint a picture of what the world would look like without peer-reviewed articles".

Dezenhall also recommended joining forces with groups that may be ideologically opposed to government-mandated projects such as PubMed Central, including organizations that have angered scientists. One suggestion was the Competitive Enterprise Institute, a conservative think-tank based in Washington DC, which has used oil-industry money to promote sceptical views on climate change. Dezenhall estimated his fee for the campaign at $300,000–500,000.

"Media massaging is not the same as intellectual debate."

In an enthusiastic e-mail sent to colleagues after the meeting, Susan Spilka, Wiley's director of corporate communications, said Dezenhall explained that publishers had acted too defensively on the free-information issue and worried too much about making precise statements. Dezenhall noted that if the other side is on the defensive, it doesn't matter if they can discredit your statements, she added: "Media massaging is not the same as intellectual debate."

Officials at the AAP would not comment to *Nature* on the details of their work with Dezenhall, or the money involved, but acknowledged that they had met him and subsequently contracted his firm to work on the issue.

"We're like any firm under siege," says Barbara Meredith, a vice-president at the organization. "It's common to hire a PR firm when you're under siege." She says the AAP needs to counter messages from groups such as the Public Library of Science (PLoS), an open-access publisher and prominent advocate of free access to information. PLoS's publicity budget stretches to television advertisements produced by North Woods Advertising of Minneapolis, a firm best known for its role in the unexpected election of former professional wrestler Jesse Ventura to the governorship of Minnesota.

The publishers' link with Dezenhall reflects how seriously they are taking recent developments on access to information. Minutes of a 2006 AAP meeting sent to *Nature* show that particular attention is being paid to PubMed Central. Since 2005, the NIH has asked all researchers that it funds to send copies of accepted papers to the archive, but only a small percentage actually do. Congress is expected to consider a bill later this year that would make submission compulsory.

Brian Crawford, a senior vice-president at the American Chemical Society and a member of the AAP executive chair, says that Dezenhall's suggestions have been refined and that the publishers have not to his knowledge sought to work with the Competitive Enterprise Institute. On the censorship message, he adds: "When any government or funding agency houses and disseminates for public consumption only the work it itself funds, that constitutes a form of selection and self-promotion of that entity's interests." ■

Jim Giles

Terror fears prompt tighter controls for UK labs

In the wake of security concerns about terrorist attacks, the UK Home Office boosted its list of 'controlled substances' used in the lab from 47 to 103 on 25 January.

University, hospital and commercial labs will have to give the government details of their exact stocks of all of these substances, which now include the virus that causes African swine fever and the strain of enterohaemorrhagic *Escherichia coli* that recently caused havoc by contaminating US spinach. If asked, they will also have to inform the police of the names of everyone handling them. The updated list also sees two new categories introduced — 18 animal pathogens and 2 fungi are now included as substances that might be of use to terrorists.

Tony McNulty, minister for police and security, says the measures are to stop terrorist groups using chemical or biological materials as terrorist weapons — a chief threat anticipated by the UK security service MI5. But some scientists say the extended list is overkill, and the increased burden of paperwork will hamper research.

Russia woos India in deal on nuclear fuel

With the much debated Indo-US nuclear deal still facing roadblocks, Russian President Vladimir Putin has offered to build four more nuclear reactors in India, in addition to the two 1,000-megawatt reactors it is already building at Kudankulam in the south of the country.

Indian officials say Russia has offered them a lifetime fuel supply and will not stop them from reprocessing the spent fuel — two crucial issues that have clouded India's deal with Washington. Russian officials say that they are still bound by the guidelines of the Nuclear Suppliers Group; but these may change and, as this regime is voluntary, it may not significantly restrict their actions.

The Indo-Russian accord, signed in New Delhi on 25 January, is seen as a signal that Russia will step in if the US deal falls through.

A. ZEMLIANICHENKO/AP

Vladimir Putin and Manmohan Singh discuss Russia's nuclear deal with India.

Fisheries lay plans to save tuna stocks from extinction

R. HERRMANN/PHOTOLIBRARY

On 26 January, after a five-day meeting in Kobe, Japan, representatives from the world's tuna fisheries issued an 'action plan' on how to save the beleaguered fish. The group agreed that urgent action was needed and decided on broad strategies, from developing catch documentation and tagging systems, to improving trade-tracking programmes and enforcing strict penalties. But as yet the plan lacks details such as numerical targets or timelines. "Their only agreement was to gather more data and talk more often," said conservation group the WWF in a press release.

Overfishing of tuna is endangering wild stocks — with some now listed as critically endangered. The number of spawning Atlantic bluefin tuna in the western Atlantic is estimated to be at 13% of 1975 levels, for example, according to a WWF report. Japan is widely blamed for the decline, as it consumes more than a quarter of the 2-million-tonne global tuna production.

The group plans to meet next in 2009 to work out more specific plans.

US set to embrace law on genetic discrimination

The US Congress looks likely to ban the use of genetic information in job-hiring and insurance-coverage decisions, after a 12-year effort by lawmakers.

The Genetic Information Nondiscrimination Act would make it illegal for health insurers to deny coverage or increase premium prices for healthy people solely on the basis of a genetic predisposition to a specific disease. It would also stop employers from using such information in making decisions about hiring, firing or promotion.

In the past, Republican House leaders have not brought the bill to a vote. But with Democrats now in charge in both houses, where the bill has bipartisan support, it looks bound for passage into law after its introduction this January. President Bush put his voice behind a ban on genetic discrimination during a visit to the National Institutes of Health earlier in the month.

Britain calls time on plan to advance clocks

A plan to bring Britain's clocks into line with those in Europe failed to gather the necessary political support and so will be dropped.

The potential change, championed by supporters as a way to save lives and energy, would have given Britain an extra hour of evening daylight all year round, by advancing the clocks by an hour from their current times. Preliminary calculations suggest that the move could have saved around 100 lives a year through reducing traffic accidents in the evenings, and about £485 million (US$950 million) in energy costs (see *Nature* 445, 344–345; 2007).

The private member's bill got marginal support in the 26 January vote, but failed to draw the necessary 100 Members of Parliament needed for the measure to proceed. Only 52 members voted, with 32 backing the bill.

Bush offers words but no action on climate change

President George W. Bush received some plaudits for referring to "the serious challenge of global climate change" in his State of the Union address last week. He called for a 20% drop in petrol use by 2017, proposed raising fuel-efficiency standards for cars, and called for more research into alternative fuels — specifically boosting investment in finding new ways to produce ethanol as a biofuel. But he did not propose any limits on carbon emissions, leading many experts to say the speech was very light on actual commitments.

In the same week, researchers at the Massachusetts Institute of Technology in Cambridge called for a renewed focus on geothermal energy as part of the country's solution to weaning itself off oil. An investment of up to $1 billion over 15 years, they said, could allow the United States to harness 10% of its electricity-generating capacity from the hot bowels of Earth by 2050. Geothermal energy currently accounts for less than 1% of US electricity use.

Correction

In our News story "PR's 'pit bull' takes on open access" (*Nature* **445,** 347; 2007), we incorrectly quoted Wiley's director of corporate communications, Susan Spilka, as writing in an e-mail: "Media massaging is not the same as intellectual debate." She actually wrote "messaging", not "massaging".

M. DI LAURO/GETTY/NEWSCOM

The US military uses a range of interrogation techniques — but do they really work?

Interrogation comes under fire

WASHINGTON DC

There is no scientific basis for current interrogation techniques, a US government-funded study has found. The report has stirred up controversy by calling for more research into the matter, angering many psychiatrists who believe such work is unethical.

The 374-page study on "educing information" was conducted by the Intelligence Science Board, an independent panel that advises the government's intelligence agencies. The report concludes that "virtually none of the interrogation techniques used by US personnel over the past half-century have been subjected to scientific or systematic enquiry or evaluation".

First published in December, the report became public last week after it was leaked to the Federation of American Scientists, a watchdog group based in Washington DC. Members of the study group declined to comment, citing the sensitive nature of their work.

The report provides a comprehensive review of military and law-enforcement interrogation techniques and finds numerous misperceptions, both within and outside professional circles. For example, it concludes that the belief that torture breaks down a subject's resistance is without technical merit, as is the effectiveness of strategies such as sleep deprivation. It also finds that professional interrogators have as many erroneous beliefs as novices about how to use body language to spot liars, and concludes that current lie-detection technology is still highly unreliable.

In a controversial final chapter, the report calls for a systematic investigation of interrogation techniques to determine which yield the best information, and suggests reviewing the testimonials of former US prisoners of war to understand whether and how torture worked on them. Finally, it calls for controlled studies on soldiers undergoing survival training and on college students willing to participate in "benign" research.

Such studies might be useful if they are conducted safely and ethically, says Steven Aftergood of the Federation of American Scientists. He points out that regardless of scientists' position on the matter, US soldiers and intelligence officers seem to be engaging in harsh interrogation practices in Iraq, Afghanistan and at Guantanamo Bay in Cuba, so they need to know what works, and what doesn't. "We have not done very well in the absence of research," he says.

Others disagree. "I doubt very much that any research could be done in a university setting or that any ethical person would do it," says Alan Stone, a psychiatrist at Harvard University in Cambridge, Massachusetts. Stone points out that interrogation is often designed to induce stress, and that raises a host of "intractable" ethical issues, such as how to gain consent from study subjects.

"We have not done very well in the absence of research."

The fields of psychology and psychiatry are split over whether to carry out such work. In 2005, the American Psychological Association stated that psychologists could participate in interrogation, but not torture. The American Psychiatric Association, meanwhile, has condemned any such work by its members. Gregg Bloche, a lawyer and psychiatrist at Georgetown University in Washington DC, says: "This underscores the need to make some rules." ■

Geoff Brumfiel

ON THE RECORD

"Of all the places to make artificial snow, this has to be the most absurd."

Jonathan Loh, co-author of a report by environmental group WWF showing that citizens of the United Arab Emirates, the desert destination for indoor snowboarding, have the largest ecological footprints in the world.

NUMBER CRUNCH

524 Nobel prizewinners and nominees from 1901 to 1950 were included in a University of Warwick survey of their average longevity.

75.8 was the average lifespan of the unlucky also-rans.

77.2 was the average longevity of the winners, suggesting that receiving a Nobel boosts more than just your career.

SCORECARD

▼ **Cigarette companies** The amount of nicotine inhaled by the average smoker increased by 11% between 1998 and 2005, research shows, because cigarettes now contain higher levels of the drug.

▲ **Quitting smoking** But smokers might be able to kick the habit more easily than before with the help of varenicline, a recently licensed drug that seems to treble the odds of successfully quitting, according to early data.

ZOO NEWS

Surprised staff at the Chimp Haven sanctuary in Shreveport, Louisiana, are to carry out a paternity test after one of the females, Teresa (pictured), unexpectedly gave birth this month — despite the males all having been given a vasectomy.

R. DELAHAYA/CHIMP HAVEN

Sources: Environmental News Network, University of Warwick, New Scientist, Cochrane Library, AP

SIDELINES

India's carbon dioxide trap

BANGALORE

Encompassing most of central and western India, the Deccan Traps is the world's largest continental flood-basalt province outside Siberia. In the 1970s, it was a candidate for storing nuclear waste, but fears of groundwater contamination killed that idea. Now it is being proposed as a site for locking away another global nuisance, carbon dioxide (CO_2) — perhaps enough to make a significant dent in global warming.

The Deccan Traps is a thick pile of solidified lava from volcanic eruptions 65 million years ago. Indian and American geologists have launched a joint study to see how well they can trap the CO_2 that has been captured from coal-fired power stations within and below the basalt layers.

Inspiration for the project came from research at the Pacific Northwest National Laboratory (PNNL) in Richland, Washington, which found that water saturated with CO_2 reacts rapidly with basalt to form stable carbonate minerals (see *Nature* 444, 792–793; 2006). Many countries have considered capturing CO_2 from power stations and storing it in aquifers or used oil wells, but there is always a risk that the dissolved gas could escape. If the gas could be converted into solid minerals within rock, however, it could be locked away long-term.

Pete McGrail of the PNNL reported the findings to a conference on CO_2 capture and storage held in Hyderabad in India on 12–13 January. The US Pacific Northwest has similar basalt deposits, which McGrail's group now estimates could hold more than 50 gigatonnes of CO_2.

"We are lucky to have the Deccan Traps, which is much bigger," says Ramakrishna Sonde, executive director of the National Thermal Power Corporation (NTPC) in New Delhi, which builds and runs India's coal-fired power stations. The NTPC is collaborating with the PNNL and the National Geophysical Research Institute (NGRI) in Hyderabad on the pilot study. Sonde says that he tentatively estimates that the Deccan Traps might be able to hold 150 gigatonnes of CO_2 — as much as the world's power industry might emit in 15 years.

The three-way project, to which India has committed US$1.3 million, is one of the 17 initiatives endorsed by the Carbon Sequestration Leadership Forum, a voluntary climate initiative of which India was a founding member. "We look at coal as a dominant energy source, so CO_2 sequestration is something we cannot ignore," Malti Goel, a senior official in the Indian science ministry and co-chair of the Hyderabad conference, told *Nature*.

DINODIA IMAGES/ALAMY

The basalt layers in India's Deccan Traps could be used to store huge amounts of carbon dioxide.

NGRI scientists admit that basalt isn't very porous, making it hard to disperse CO_2 through the rock, but say that this is compensated by its high reactivity with the gas. The ultimate idea is to pump supercritical CO_2 into porous sedimentary rocks below the basalt layer. The gas would move upwards though the rock and react with the basalt above, forming a 'cap' that would stop any unreacted gas from escaping.

"We look at coal as a dominant energy source, so CO_2 sequestration is something we cannot ignore."

Some geologists, however, seem to be sceptical about the whole project. "I will not recommend CO_2 sequestration in volcanic areas," says Kiran Shankar Misra, deputy director general of the Geological Survey of India. "These are regions that are hot, and the basalts are highly fractured." He says that he is worried that the heat in the rock would cause the dissolved CO_2 to become gaseous, and that it could then seep out through the fractures. "The CO_2 will not stay there."

Hetu Sheth, a geologist at the Indian Institute of Technology, Bombay, points out that unlike the Columbia River basalts, which are only 14 million to 17 million years old, the Deccan lavas are highly weathered, meaning that they will already have reacted with CO_2 in the air to a certain extent. "Their capacity for reacting with newly injected CO_2 may therefore be low, certainly lower than the Columbia River basalts," says Sheth.

Before any CO_2 is injected into the Deccan basalts, an initial phase of the project, scheduled to take 18 months, will characterize the permeability and porosity of the rock. It will also look at the nature of faults within the rock, to determine how CO_2 would flow through the rock once injected. The PNNL is due to start pumping CO_2 into the Columbia River basin basalts later this year, a project that Sonde says India is "keenly watching". ■

K. S. Jayaraman

A. JAREKJI/ REUTERS

CROWD RESEARCH MAKES PILGRIMAGE SAFER
The annual Hajj to Mecca gets an update based on the science of pedestrian motion.
www.nature.com/news

Rebels hold their own in journal price war

Last August, the entire editorial board of the Elsevier journal *Topology* quit in a row over pricing. Now they are setting up a non-profit competitor to be published by the London Mathematical Society. The *Journal of Topology*, announced last week, will launch next January and will cost US$570 per year, compared with *Topology*'s $1,665.

It's not the first such move. Over the past eight years, around a dozen cheap or open-access journals have been created to compete directly with an expensive commercial journal, many by editorial boards that had quit the original publication in protest. So, do the cheaper journals fare better than their rivals?

As far as scientific credibility is concerned, the answer is often yes — many of the challengers have obtained impact factors (a measure of the citations its papers receive) higher than their competitor. For example, the *Journal of Machine Learning Research*, set up in 2001 by editors of Springer's *Machine Learning*, has a 2005 impact factor of 4.027. "That's the highest in artificial intelligence, automation and control, and ninth in all of computer science," says Leslie Pack Kaelbling, a researcher at the Massachusetts Institute of Technology in Cambridge and the journal's editor-in-chief. *Machine Learning* has a 2005 impact factor of just 3.1.

One of the first defections was *Evolutionary Ecology Research*, set up in 1999 by Michael Rosenzweig, a researcher at the University of Arizona in Tucson. He defected, along with the entire editorial board, from *Evolutionary Ecology*, then published by Wolters Kluwer. Rosenzweig had established the journal in 1987, but became disillusioned with price increases.

"Despite scholarly success, the rebel journals often get poor support from libraries."

The new journal has survived, with the online version getting up to 80,000 page views a month and a 50% increase in its number of pages, says Rosenzweig. He says the journal pioneered the idea of allowing libraries to subscribe only to an electronic version. Although its impact factor of 1.61 lags behind *Evolutionary Ecology*'s 1.77, he dismisses this as unimportant: "Impact factors are toxic for science and the journals that serve it."

Like many of the rebel journals, *Evolutionary Ecology Research* was set up with help from the Scholarly Publishing and Academic Resources Coalition (SPARC), which fosters greater access to the literature. Joan Birman of Columbia University in New York is on the editorial board of *Geometry and Topology*, another SPARC-supported journal. She says it "has done better than we could have dreamed".

But despite scholarly success, the journals often get poor support from libraries. "Libraries have not given us anything like the support, via subscriptions, that mathematicians have given us via submissions," says Birman. "Subscriptions have been an uphill battle. Our journals are self-supporting, but just barely so." ■

Declan Butler

See www.nature.com/news/2007/070122/full/445351a.html for a table of rebel journal success.

Demolished satellite turns into dangerous debris

One unwelcome outcome of China's recent destruction of a satellite is the creation of a large amount of space debris.

On 11 January, according to US intelligence sources, China launched a test weapon that struck and destroyed an obsolete weather satellite.

David Wright, a weapons expert at the Union of Concerned Scientists in Cambridge, Massachusetts, estimates that the explosion created a cloud of 2 million particles in an orbit similar to that of many other satellites. The debris probably includes some 40,000 particles larger than 1 centimetre moving at about 7.5 kilometres a second, 30 times the speed of a jumbo jet.

"A millimetre-sized piece of debris can very seriously damage a satellite," says Wright. Many of the particles will remain in orbit for a decade or longer, he adds.

Drop in fatalities fuels optimism over cancer

The number of people dying from cancer in the United States has fallen for the second time in two years. The second decline has convinced experts that the drop — a first since records began in 1930 — marks a real trend rather than a statistical fluke.

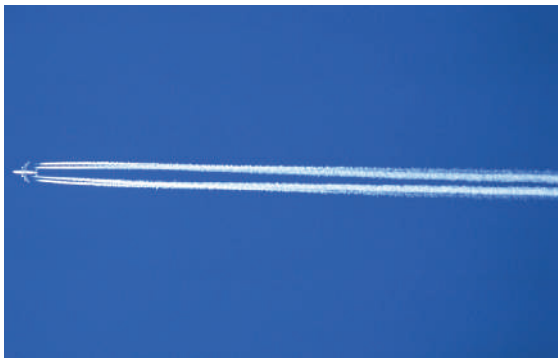
The latest figures were announced by the American Cancer Society on 17 January (A. Jemal *et al. CA Cancer J. Clin.* 57, 43–66; 2007). Compiled for 2004, they show that the number of deaths fell by 0.5% over the previous year to 553,888. Earlier tumour detection, better treatment and disease prevention due to lifestyle changes are probably behind the drop, the society says.

But the decline doesn't mean doctors and researchers can drop their guard. Cancer still kills more than 7.6 million people each year, accounting for about 13% of all deaths worldwide.

Britain draws up rules for carbon buy-back schemes

Britain is set to become the first country to introduce government-regulated voluntary standards for carbon-offsetting schemes. The move would allow consumers to ensure that money paid into such projects goes towards cutting greenhouse-gas emissions.

Under the scheme, offsetting agencies that measure up to the government's code of practice will get a 'quality mark' that they can advertise to customers, said environment minister Ben Bradshaw, who opened the plans up to public consultation on 18 January.



Pay as you go: a UK scheme could boost consumer confidence in ways to offset carbon emissions.

F. DAMM/ZEFA/CORBIS

Carbon offsetting (see *Nature* 444, 976–977; 2006) gives consumers the chance to buy back greenhouse-gas emissions from activities such as flying by investing in emissions-saving projects. Some businesses have begun offering offsets as part of holiday packages, but consumer groups have warned that not all schemes deliver genuine emissions cuts. The new scheme would give a thumbs-up to agencies that offer offsetting schemes approved by monitoring bodies such as the United Nations' Clean Development Mechanism.

Growth of transgenic crops seeds food fight

For the first time, the area of the world planted with genetically modified crops has exceeded 100 million hectares, according to an industry-backed group. The figure represents a 13% jump over 2005, it says, and shows a 60-fold increase since the crops were first planted a decade ago.

In its report issued on 18 January, the International Service for the Acquisition of Agri-biotech Applications notes that more than 90% of the 10.3 million farms growing biotech crops are relatively small.

But opponents of genetically modified crops published their own report on 9 January asserting that uptake of the technology has generally increased the use of pesticides and has not benefited either small farms or consumers.

The Amsterdam-based Friends of the Earth International and the Center for Food Safety, based in Washington DC, say that most genetically modified crops are used as high-priced animal feed to supply rich nations with meat. They report that more than 70% of the crops are grown in the United States and Argentina, and claim that genetically modified crops have served to boost herbicide sales while increasing the number of herbicide-resistant weeds.

Dutch astronomer shines at European observatory

Astronomer Tim de Zeeuw has been appointed director-general of the Munich-based European Southern Observatory (ESO). When he takes office in September, he will become the observatory's seventh director — and the fourth to be Dutch. At 50 years old, he is also one of the youngest to hold the post.

Currently science director of the Leiden Observatory in the Netherlands, de Zeeuw will temporarily exchange his personal research interests in galaxy and star formation for the technical, organizational and financial management of the ESO's three major projects: an upgrade of the Very Large Telescope in Chile, construction of the Atacama Large Millimeter Array, and development of the European Extremely Large Telescope.

ESO

Starman: Tim de Zeeuw will take the reins at the European Southern Observatory in September.

Think-tank highlights rapid rise of science in Asia

Surges in markets, state funding and a flow of native talent heading home are boosting science and innovation in China, India and South Korea to an unprecedented extent that is too little appreciated elsewhere. That's the main conclusion of four reports surveying these countries, published last week by the UK think-tank Demos.

The reports also note the countries' weaknesses, including a need to establish more confidence in ethical frameworks and to develop home-grown creativity. Nevertheless, Demos says, a fundamental shift in the geography of scientific ideas and their impacts is under way.

Demos warns governments in the West not to react in a defensive way to this growth. Instead the group proposes schemes by which Britain, in particular, should increase its engagement with the three countries.

BUSINESS

Crunch time for multiple-gene tests

Sophisticated new genetic tests face an uncertain future — unless they can win clear-cut approval from regulators, insurers and, most importantly, doctors. **Virginia Gewin** reports.

For many of the women who will learn this year that they have early-stage breast cancer, chemotherapy won't do any good. And a sophisticated genetic test is already available that might help to predict who they are.

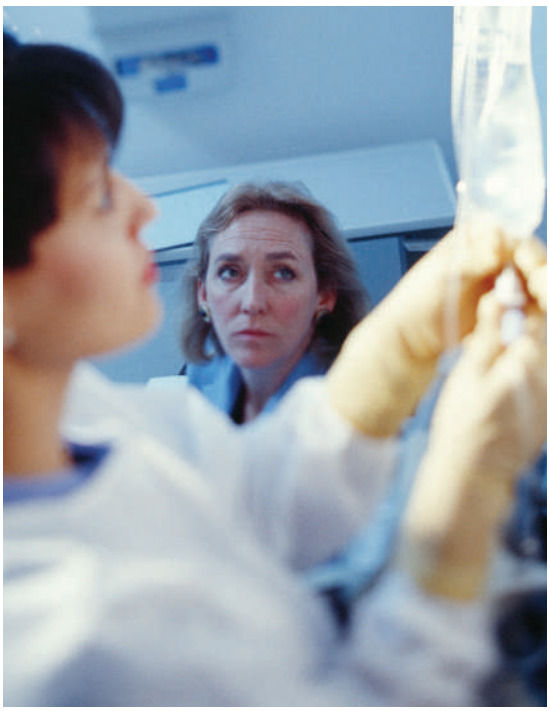
Some say that the test — Oncotype DX, developed by Genomic Health of Redwood City in California — is a portent for the long-awaited and much-hyped era of 'personalized medicine', in which therapies are fine-tuned to meet patient needs.

The 21-gene screen, which costs about US$3,400 per test, uses an algorithm to generate an individualized disease 'recurrence score', which indicates which patients are most likely to benefit from chemotherapy.

But such diagnostic tests have, so far, been subject to uneven regulatory review. Of the thousand-or-so genetic tests for specific diseases developed in the past decade in the United States, most are regarded as laboratory-developed, or 'home brew' tests. For these, the laboratories, not the tests, are subject to oversight by the US Centers of Medicare and Medicaid Services. Only six home-brew tests have so far been designated as medical devices by the US Food and Drug Administration (FDA), and therefore subject to regulatory approval. "We believe that those tests that meet the criteria of medical devices are subject to FDA regulations, but because of limited resources, we have applied our enforcement discretion," says Steve Gutman, the FDA official in charge of their evaluation.

The FDA is now inviting public comment — due by 5 March — on how it should regulate tests that use proprietary algorithms to predict risk. The future of such approaches in the United States, where their use is most widespread, could largely hinge on what it decides to do. Until the decision is made, however, the use of the tests is spreading as insurance companies — which make the critical decisions in how US healthcare is administered — decide to pay for them.

As diagnostic companies craft more complex tests that combine data from genome, proteome and even cellular metabolites to predict disease risk and recurrence and to select therapies, they have garnered increasing venture-capital support. And Genomic Health has seen its stock grow steadily in value since it made its debut on the Nasdaq in September 2005 (see graph).

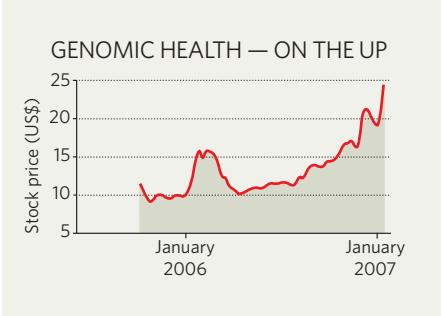


W. & D. MCINTYRE/SPL

Ideally, complex genetic tests could aid patients facing tough decisions over cancer treatment.

The FDA review will now clarify rules for existing tests and map out suggested regulations for the new ones such as Oncotype DX. These tests are collectively known as *in vitro* diagnostic multivariate index assays, and use multiple tests and sophisticated algorithms to generate results.

If the FDA decides to play tough, it could insist on 'prospective studies' — ones that follow up patients' progress for several years — before approving the tests. But the cost of doing these trials would be prohibitive, says Vijay Aggarwal, chief executive of Aureon Laboratories in Yonkers, New York.



Last February, Aureon launched its Prostate Px test, which combines molecular markers and advanced imaging to predict prostate cancer recurrence after the prostate has been removed. Aureon has validated the test to its own satisfaction by studying the tissue of patients who were identified retrospectively over five years. Aggarwal says that an FDA requirement to do prospective testing instead, for a disease that can take years to manifest itself, could stifle diagnostic companies such as Aureon. "A prospective trial in prostate cancer could take 20 years to complete," he admits. "No company has the funding for that."

Genomic Health's own validation efforts have raised the bar for the entire industry: it says that it spent $100 million to validate its test results through three separate clinical trials. This may have set a standard that few other diagnostics companies will be able to match, says Sean Tunis, head of the Center for Medical Technology Policy, a consultancy based in San Francisco in California.

But even if a test can be shown to work, its success will ultimately depend on the extent to which it affects treatment decisions. "There is no reason to do the test, or have an insurer pay for it, if the doctor doesn't take its advice," says Lee Newcomer, chief medical officer at Minneapolis-based United HealthCare, one of the largest US health insurers. Although United recently agreed to cover Oncotype DX, the company will not cover it when patients who are identified as low-risk by the test are subsequently treated with chemotherapy.

So far, Genomic Health has managed to secure payment for the Oncotype DX tests from insurers that cover about 40% of the US healthcare market — including Medicare, the government system for the elderly. "The absence of a regulatory environment doesn't mean the marketplace is not forming," says Rick Carlson, a health policy specialist at the University of Washington in Seattle.

And as they seek wider adoption of their tests, diagnostics companies have another

factor to consider: whether they will be helped or hindered by jumping into bed with the drug companies.

Some argue that the best way to ensure market success is to link the diagnostic test to a specific drug treatment. But big pharmaceutical companies have been reluctant to collaborate with the diagnostics companies — fearing, according to some observers, that markets for their drugs could only be shrunk by effective tests.

But now major drug companies are showing more interest. Last May, for example, Pfizer invested $25 million in a diagnostic test developed by Monogram Biosciences of San Francisco to identify patients who are likely to respond favourably to their next-generation HIV blocker drug. Others, including Eli Lilly and Merck, have partnered with diagnostic companies to develop tests that can identify patients who will benefit from their therapies.

"Having a relationship with pharma certainly creates a tremendous opportunity for a diagnostic company," says Aggarwal, whose company is also partnering with Pfizer. "There is a significant market opportunity for us to predict outcomes for post-biopsy prostate cancer. But the real goal is to link our diagnostic test to a therapeutic intervention."

"There is no reason to do the test, or have an insurer pay for it, if the doctor doesn't take its advice."
— Lee Newcomer

And according to Peter Keeling, whose London-based consultancy firm, Diaceutics, specializes in liaison between pharmaceutical and diagnostics companies, the latter group do need such links. He says that Genomic Health's model is too expensive to serve as a role model for most other diagnostics firms.

Randy Scott, chief executive of Genomic Health, says that the company will work with pharmaceutical partners, but doesn't want to be reliant on them. "It's more critical right now to develop a strong independent diagnostics industry, doing high-quality pharmacogenomics work, and then to work our way back to the early stages of drug development," he says.

But according to Carlson, only a handful of tests are in the pipeline that have enough evidence behind them to affect clinical decision-making. He says that the diagnostics industry is caught in a catch-22 situation: their products cannot generate the revenues to pay for the testing that would prove their utility. "Although few doubt that these technologies will make a difference over time," he says, "the industry is now in the trough between hype and hope." ■

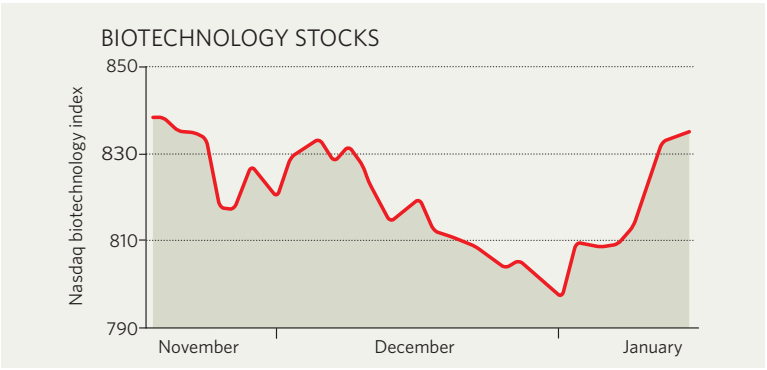
IN BRIEF

DRUG DRAW European regulators are approving new drugs just as quickly, on average, as the US Food and Drug Administration (FDA), according to an assessment by the Tufts Center for the Study of Drug Development in Boston, Massachusetts. Tufts analysed the period from 2000 to 2005 and found that of 71 new drugs approved by both the European Medicines Agency and the FDA, the average times to approval were 15.8 months and 15.7 months, respectively. The FDA approved most of the products — 47 of them — more quickly than the European regulator, but its approval times were much more variable.

WOOD RECYCLING A Japanese company will next month start operations of a ¥4-billion (US$33-million) facility to produce ethanol from scrap timber. The plant will consume up to 50,000 tonnes of timber in its first year, generating about 2 megawatts of electricity and 1,400 tonnes of bioethanol for use as an additive to petrol. It has been built by Bio Ethanol Japan Kansai in Osaka, based partly on technology developed by Celunol in Massachusetts. Corn and sugar cane have traditionally been the main feedstocks for ethanol plants.

STALLING TACTIC The US Federal Trade Commission (FTC) says that pharmaceutical companies are successfully making use of a new tactic aimed at delaying the introduction of cheaper, generic drugs. The commission told a US Senate committee on 17 January that 14 patent litigation settlements between drug firms and generic companies in the fiscal year ending 30 September included 'pay-for-delay' agreements. In these, a generic competitor is paid by the maker of the patented drug to delay introduction of a copycat drug. The previous year, the FTC had counted three such agreements, and there were none the year before that.

MARKET WATCH



SOURCE: NASDAQ

This week, Wood Mackenzie, an Edinburgh-based research and consulting firm, reviews recent trends in biotechnology stocks.

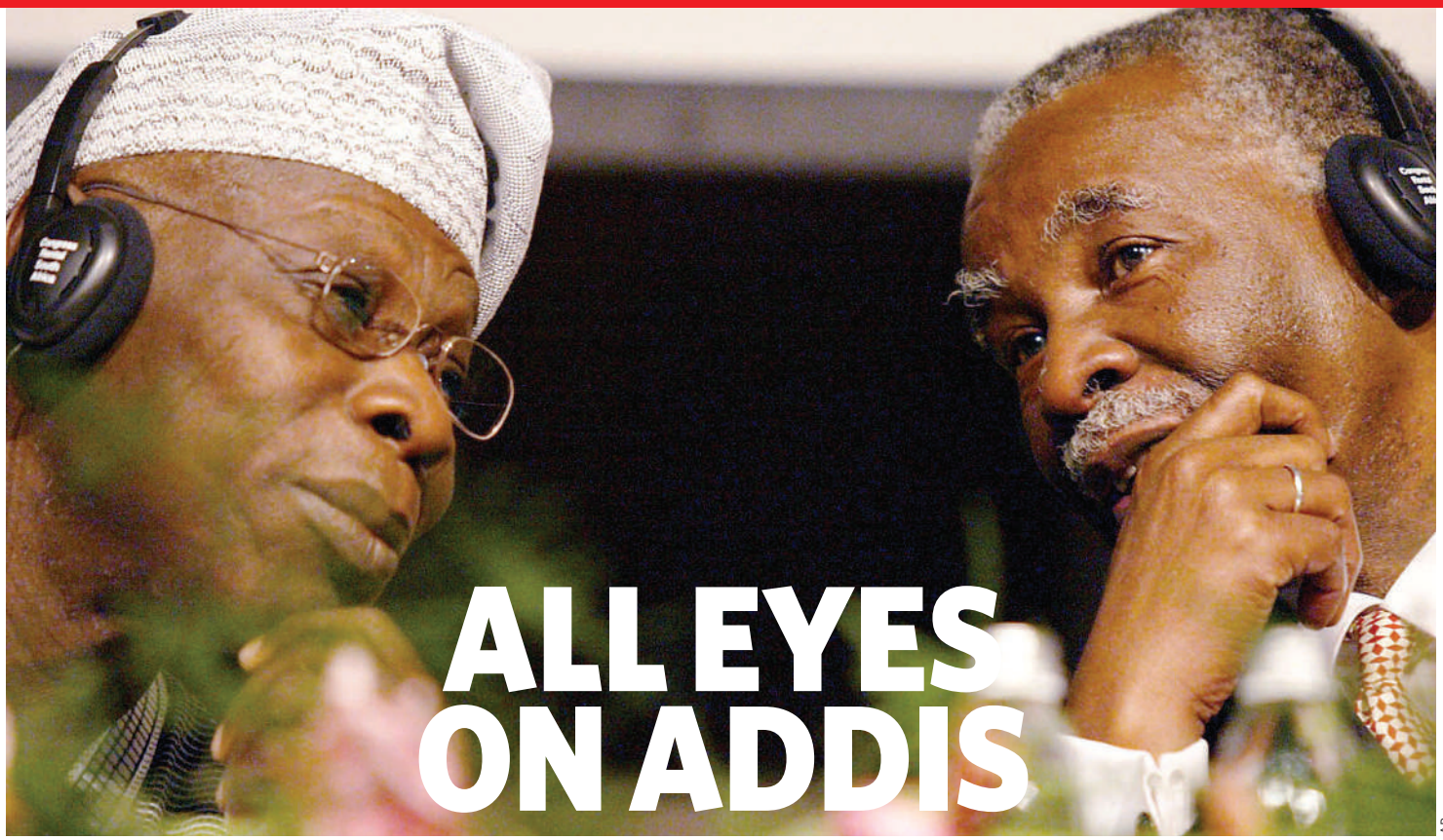
The Nasdaq Biotechnology Index fell steadily as 2006 drew to a close, but rebounded in early January to its mid-November levels.

Downturns in individual stocks were precipitated by poor data from clinical trials. Shares in Illinois-based Neopharma, for example, plunged by two-thirds on 8 December after news that its oncology therapy cintredekin had not shown a survival benefit in aggressive brain tumours. On the same day, Nuvelo of California met a similar fate, with its stock losing four-fifths of its value when its anti-clotting candidate, alfimeprase, failed to help patients with poor peripheral circulation in two trials. That could also jeopardize Nuvelo's lucrative co-development agreement with Bayer in Germany.

Later gains were bolstered by the announcement of partnership and licensing deals. Shares in Epix Pharmaceuticals of Massachusetts rose by 30% in mid-December, after news of its drug development collaboration with GlaxoSmithKline on a potential treatment for Alzheimer's disease.

And Seattle Genetics climbed by 34% in early January, after licensing commercial rights to its early-stage anticancer monoclonal antibody, SGN-40, to Californian industry leader, Genentech.

Overall, the index is emerging from an indifferent 2006: it made impressive gains at first to end February up by 21%, its high point for the year, only to slide by mid-July to 12% below where it started the year. A subsequent recovery took it to an anaemic 1% gain for all of 2006, underperforming other, broader market indices. ■



ALL EYES ON ADDIS

AP

Next week, African leaders will come together to talk about science and technology at a summit in Ethiopia. This presents an opportunity to allot some foreign aid and, if they get it right, to launch projects that will draw further donations from abroad, says **Michael Cherry**.

If asked to name Africa's highest priorities, most people would cite poverty, disease and conflict, not science and technology. But at next week's summit of African leaders in the ancient Ethiopian capital of Addis Ababa, science and technology research for African development is top of the agenda. "The challenge," says South African director-general of science and technology Phil Mjwara, "is to impress upon the heads of state that science and technology is of such critical importance to the continent's development that it should be a policy priority for every African nation."

For many, this meeting of the five-year-old African Union (AU) offers a historic opportunity. David King, chief scientific adviser to the British government, argues that, apart from boosting the economy and creating employment opportunities, "the important point about investing in science and technology is that it raises the level of aspiration throughout the educational system." Indeed, some now view science and technology as a vehicle for spending foreign aid wisely.

Eyes in the developed world as well as in Africa will be watching the summit keenly, especially because the new United Nations secretary-general, Ban Ki-moon, is scheduled to put in an appearance. Africa's leaders could use this summit as an opportunity both to convey their own financial commitment to the development of science and technology, and to begin work on detailed plans that foreign donors can support.

During the past 18 months the world's rich nations have expressed renewed interest in addressing Africa's fate. In July 2005, the Gleneagles summit committed the G8 countries both to eliminating debt and to providing a significant injection of aid. This renewed interest in Africa is accompanied by the realization that past aid was often squandered, either on corrupt officials in developing countries or on foreign consultants. The latter problem can be addressed by removing restrictions on where aid is spent; spending new money wisely is a challenge that Africa must meet, and the hope is that science and technology offers one route to doing so.

"I will consider the summit a success if a handful of leaders return home emboldened to champion the role of technological innovation." — Calestous Juma

The AU has been gearing up to this summit over the past two years, charging its ministerial council on science and technology (AMCOST) to come up with a wish list of appropriate research and development for the continent — the Consolidated Plan of Action. The list of research clusters covers 12 flagship projects, and with the exception of space science, all are targeted at fairly obvious African concerns (see 'Africa's 12-point plan'). Surprisingly, biomedical science is omitted, but according to Botlhale Tema, the AU's director of human resources, science and technology, this is not because the AU is unconcerned about research on human disease, but because it falls under a different department (social affairs) of the AU. Mjwara is positive about the plan: "All of the proposed projects are in fields in which there is already existing momentum based on at least some local expertise."

The more difficult issue of implementation is largely unresolved. A small meeting of African scientists in Alexandria, Egypt, last October produced some very vague recommendations, and a further meeting of science ministers in Cairo in November made some suggestions for implementing the Consolidated Plan of Action, embodied in the Cairo declaration. "More important than the actual recommendations that came out of either meeting is the fact that they took place at all," says Andy Cherry, science and technology adviser at the Association of Commonwealth Universities in London, who attended both meetings as an observer.

The Cairo declaration, which is likely to be ratified by the leaders at the Addis summit, contains recommendations on potential funding mechanisms and 'centres of excellence' for

moving the plan forward, as well as possible agreements on biotechnology and biosafety. Biotechnology is one area on which the summit could conceivably reach agreement, because a fairly detailed report on both priorities for biotechnology research, and biosafety measures for genetically modified (GM) organisms, will be on the table.

Increasing integration

Kenyan Calestous Juma of Harvard University, a member of the panel that compiled the report, hopes that biotechnology development and regulation can proceed hand-in-hand. He believes the initial focus should be on making effective use of products that are relevant to local needs and are ready for commercialization, including techniques for disease control, pest tolerance and weed management. "I will consider the summit a success if a handful of leaders return home emboldened to champion the role of technological innovation in developing their country and region," he says.

The report also encourages African nations to abandon their individual policies on GM crops in favour of a common, or at least regional, stance. South Africa and Kenya, for example, have promoted the introduction of GM crops such as corn (maize), cotton and soya, whereas Zambia has banned all such crops. Until 2004, the European Union had a five-year moratorium on GM imports, and getting approval for new products is still difficult and time-consuming. As Europe is the main importer of African agricultural products, and because most African nations lack facilities for separating GM and normal crops, many have been reluctant to pursue the technology.

Because Africa is much less integrated than the European Union, with economic integration existing at a regional, but not a continental, level, Juma thinks regional consensus on biotechnology and biosafety is most likely. "If biotechnology development and regulation are to go hand in hand, then the AU needs to develop a strategy broad enough to empower regional economic communities to take differing positions," he says.

Fund management

On the thorny issue of funding mechanisms, the Cairo declaration backed the establishment of an African Science and Innovation fund to take forward the research themes proposed in the Consolidated Plan of Action. As an intergovernmental entity, under the auspices of the AU, the fund would support five of the recommended 12 flagship research projects in its first two years of operation, and eventually have 12 running at any given time.

"Science and technology is of such critical importance that it should be a policy priority for every African nation" — Phil Mjwara

But the Cairo meeting was unable, despite intense debate, to make recommendations about either the governance or the administration of the facility, apart from deciding that it should be based on existing organizations rather than an entirely new one. One option would be to contract fund management to the African Development Bank. But it seems unlikely that such details will be decided and agreed in Addis.

The most that can be expected at the summit would be pledges towards initial funding from individual African governments. An initial AU estimate for funding the Consolidated Plan of Action over the next 5 years asks for US$158 million. One AU-backed proposal suggests that funding sources should include African governments, foreign donors — including bilateral and international aid agencies and foundations — and private-sector contributions from Africa and elsewhere.

Africa's 12-point plan

Africa's Science and Technology Consolidated Plan of Action is made up of research clusters covering 12 projects.

- Conservation and sustainable use of biodiversity
- Safe development and application of biotechnology
- Securing and using Africa's indigenous knowledge base
- Building a sustainable energy base
- Securing and sustaining water
- Combating drought and desertification
- Building Africa's capacity for material sciences
- Building engineering capacity for manufacturing
- Strengthening the African Laser Centre
- Technologies to reduce post-harvest food loss
- Information and communication technologies
- Establishing the African Institute of Space Science

Khotso Mokhele, former president of the South African National Research Foundation, the continent's largest funding agency, argues that African states need to start by making a financial commitment themselves. "If an agency or fund is established simply as a vehicle for handouts by the West, it is doomed to failure," he says. ALC

Similar pledges have been made before by African science ministers, but they need to be supported by heads of state. In 2003, at the launch of AMCOST, African states pledged to work towards a spending target of 1% of gross domestic product on science and technology, compared with a global average of 2.36%. At the time, Egypt was the only African state reaching this target, with Algeria and Uganda coming close. South Africa announced in December that anticipated increases should allow it to reach the 1% target by 2008. But most other nations are still some way off.

High hopes

Another element of the Cairo declaration requiring additional funding is the establishment of centres of excellence to take forward some of the research in the Consolidated Plan of Action. Proponents of the concept point to India's four large science and technology institutes, which arguably provided the foundation for the country's sound base in science and technology. Britain's Commission for Africa proposed last year that the international community provide up to $3 billion over the next ten years to develop such centres.

The problem with establishing centres of excellence is that, by definition, funding becomes localized in certain areas in preference to others. This makes any decisions about which institutes to build and fund very political. In certain scientific fields Africa will have to create new institutes, whereas in others existing ones can be expanded. In the area of biotechnology (which has been broadly defined), for example, AMCOST has recommended building on existing institutions, or regional networks of institutes.

Not surprisingly, these reflect regional strengths and priorities: pharmaceutical biotechnology would be the focus in north Africa, with other healthcare biotechnology (including medical diagnostic testing kits, and stem-cell research) in southern Africa. In east Africa, biotechnology research on breeding and feed for livestock would build on expertise at the International Institute for Livestock Research (ILRI) in Nairobi, Kenya, and, in west Africa, biotechnology for crop improvement would extend the work of the West Africa Rice Development Association. Finally, central Africa would target biotechnology for forest

conservation and development — one possibility being the use of DNA barcoding to conduct taxonomic studies in forests.

On the ground, increased funding would allow these institutions to extend their activities. For example, John McDermott, director of research at ILRI, says it would allow his institute to make its state-of-the-art facilities in genomics and immunology available to other researchers in the region. McDermott wants ILRI to be a regional platform for biotechnology. "This is important because capacity in science and technology is directly linked to capacity for innovation, which the African economy so urgently requires," he says.

The big picture

But DNA barcoding in forestry research seems far removed from the daily concerns of most ordinary Africans. One criticism African leaders face whenever they focus on science and technology is that it is an élitist concern, distant from the realities of poverty, disease and infrastructure in Africa. Complaints about low representation at the pre-summit meetings in Alexandria and Cairo only strengthen such concerns. The Alexandria meeting in particular was billed as an opportunity for African scientists to provide input, yet it was attended by only 110 African delegates, of whom less than one fifth were active researchers.

Similar charges of élitism have been levelled at the African Institutes of Science and Technology (AIST) initiative, a possible model for African centres of excellence. First proposed by former World Bank president James Wolfensohn, the initiative hopes to establish four élite universities in science and technology in sub-Saharan Africa: one each in the north, west, east and south. Its proponents argue that it addresses crucial manpower needs by offering first degrees in science and engineering (including some business modules), as well as postgraduate degrees. The institutes will be overseen by the Nelson Mandela Institute, a non-profit company registered in Delaware, the board of which includes former presidents Nelson Mandela of South Africa and Joachim Chissano of Mozambique.

G. PIROZZI/PANOS

GM crops could be more widely grown across Africa if consensus can be reached on their use.

The first institute will be founded later this year in Abuja, Nigeria, although a president has yet to be appointed. It has substantial support from the Nigerian government, and plans to enrol its first students in September next year. At full capacity, it is expected to have 5,500 students, 40% of whom will be in graduate programmes. Further institutes are planned in Burkina Faso and Tanzania.

Wole Soboyejo, a Nigerian materials scientist at Princeton University who chairs AIST's African Science Committee, rejects the idea that the concept is élitist. He argues that enhancing manpower capability is critical: sub-Saharan Africa, for example, produces only about 83 engineers per million of the population annually, whereas, on average, developed countries produce 1,000. He emphasizes that "the institutes will not be islands of excellence, but will be linked to existing research centres and universities in each region, for example, providing them with access to electronic journals". Soboyejo suggests that closer cooperation with the AU would also be desirable.

Mjwara is concerned that the new institutes "might not be sustainable as they do not seem to be based on local capacity". He would prefer the AU's proposed centres of excellence to follow the example of the Pretoria-based African Laser Centre, with its network of six national facilities in Tunisia, Senegal, Ghana, Algeria, Egypt and South Africa.

Soboyejo, however, defends the model proposed for AIST by comparing it to another South African success story, the African Institute for Mathematical Sciences (AIMS) in Muizenberg. Like AIMS, the new institutes would combine local expertise with that of foreign academics teaching short courses. "At present, there is huge goodwill towards Africa worldwide, and the challenge is to channel this into returning lasting benefits for the continent," he says.

> **"The institutes will not be islands of excellence, but will be linked to existing research centres and universities."**
> — Wole Soboyejo

This call is echoed by Juma, who will deliver a keynote address at the summit. "The debate urgently needs to shift from calls for funding to thinking about creative ways of using existing scientific and technical knowledge to solve local problems," he says. And in so doing, Juma believes the linkages between institutions of higher learning and research will need to be re-examined. "More attention should be paid to rebuilding these institutions as vehicles of community development," he explains. Many African universities operate like those in the West did 40 years ago, with few links to either industry or local communities.

King is optimistic about the summits' chances of reaching agreement on the infrastructure required to go forward: "As soon as we have a realistic set of proposals on the table, we will be in a position to discuss how to fund them." Concrete policy proposals, backed by foreign investment, would surely be welcomed by African scientists. "We've had a lot of rhetoric so far about the potential importance of science and technology in African development," says McDermott, "but not much action." ■

Michael Cherry is *Nature*'s contributing correspondent in South Africa.

See Editorial, page 339.

J. SILBERBERG/PANOS

The development of African centres of excellence could see the creation of new institutions and the expansion of existing ones.

B. MELLOR

IT'S ALL IN THE TIMING

Taking hormones to replace those lost during menopause helps many women with their symptoms, yet it may also cause cognitive decline. Could the age at which hormones are taken determine whether they will be beneficial or harmful? **Tom Siegfried** reports.

Ageing female brains are perplexing, and not just to ageing husbands. Male and female neuroscientists are confused about the role that sex hormones play in mental decline after menopause.

Fuzzy memory and other mental maladies often accompany menopause, when natural production of steroidal sex hormones — mostly oestrogens — diminishes. Until a few years ago, evidence suggested that replacement of the lost hormones could alleviate memory problems and even protect against dementia later in life.

But in 2003, a large clinical trial performed in the United States reported that a popular type of hormone replacement therapy (HRT) did not improve cognition — and actually raised the risk of dementia[1,2]. That result baffled many scientists, as it contradicted both animal experiments and preliminary human studies showing that HRT benefited the brain.

But a better understanding of the biology that governs the brain's response to HRT might help to explain these discrepancies. The latest research is revealing a neurochemical soap opera about the brain's relationship with the sex organs, and how that relationship turns turbulent as the body ages[3].

The latest twist in the plot suggests that HRT can indeed aid the brain, especially if the timing is right. In fact, when given to a relatively young and healthy brain, HRT could well protect it against later damage. But when given to a brain that is already in decline, it can make things worse. "Timing is everything," says Andrea Gore, a neuroendocrinologist at the University of Texas at Austin.

After all, the landmark study that indicated possible mental declines under HRT — part of the massive Women's Health Initiative — tested women aged 65 to 79. That is long past the average age of menopause in the Western world, at about 51 years old. Women who start HRT around the time of menopause, however, might instead reduce their risk of dementia, says Peter Schmidt of the National Institute of Mental Health in Bethesda, Maryland, who studies the brain's response to sex hormones.

> **"The time and the age at which a person takes hormone therapy may predict the clinical outcome."**
> — Peter Schmidt

"There may be a critical window within which, if oestrogen is started, it will have a beneficial effect on the brain," he said in October in Atlanta at a meeting of the Society for Neuroscience. "The time and age at which a person takes hormone therapy may predict the clinical outcome."

Nobody knows for sure whether such a window exists, or precisely when it would be open. And even if the benefits to the brain are confirmed, the Women's Health Initiative study identified other hazards from HRT, such as an elevated risk of breast cancer. The findings drove a dramatic drop in the number of postmenopausal women using HRT. In 2001, by one estimate, about 15 million women in the United States received HRT; that number fell to about 10 million after the trial reported on its dangers[4].

Women are now generally advised to take HRT for as short a time as possible (if at all) and only for treatment of menopausal symptoms. So, for HRT to help the brain, new hormone replacement molecules will probably be needed that confer the benefits without the risks.

To help in the hunt for new therapeutic molecules, researchers have realized that a deeper understanding is needed of the brain's role in menopause. For decades, research into menopause focused on the ovaries, which start to lose their hormone-producing follicles at age 40–50. "It really puzzled me that there was very little interest in whether the brain may also play a primary role in the control of reproductive ageing," says Gore.

But research in Gore's and several other laboratories has shown that menopause does not simply result from the ovaries' decline. It also encompasses complex cross-talk between these organs and the brain. Menopause occurs in stages, and hormone secretion fluctuates considerably during the 'menopause transition' — the time from the first changes in the activity of the ovaries to the year after the final menstrual period.

That hormone variability reflects the shifting action in a three-character play starring the hypothalamus, a small structure at the base of the brain (see graphic). Supporting actors include the pituitary gland and the gonads (in women, the ovaries); collectively, the three make up the hypothalamic–pituitary–gonadal (HPG) axis.

The activity along the HPG axis is thought to be driven by the hypothalamus. Neurons there produce a peptide hormone known as gonadotropin-releasing hormone, or GnRH, which travels to the nearby pituitary gland and stimulates it to secrete other hormones. In turn, these hormones circulate to the ovaries, which respond to this chemical chain-mail by releasing their own hormones that travel back to the brain.

In younger adults, the characters in the HPG axis play their parts robustly, communicating clearly to maintain the body's reproductive ability. Reproductive ability declines once the hypothalamus starts to falter in how it responds to feedback from the ovaries — especially to signals sent by the main type of oestrogen, known as oestradiol.

Several actors conspire to twist the plot at this point. Oestradiol messages cannot signal the GnRH neurons directly, but instead stimulate a type of receptor on neighbouring neurons in order to relay the signal. But that nearby receptor starts to exit the stage as the body ages. "The expression of that receptor changes between young and middle-aged animals," Gore notes.

Early retirement

The GnRH neurons themselves remain on-stage, but no longer respond as effectively to the signals that do get relayed from their neighbours. The problem, Gore says, seems to be related to subtle alterations in receptors for glutamate — the courier for the signals — on the ageing GnRH neurons.

Other players in the drama also diverge from the original script. Chemical cues sent to the hypothalamus from the ovaries change with age. In fact, all three levels of the HPG axis are changed in ageing rats and monkeys. "The responsiveness of this hypothalamic network to ovarian input decreases with ageing," says Gore. "Here is where timing is everything."

For some reason, she says, the hypothalamus gets old while the rest of the brain remains middle-aged. That hypothalamic senescence can trigger menopause-related mental decline years before deficits in learning and memory would normally start to become evident.

When the hypothalamus loses interest in the messages from the ovaries, hormone production diminishes. But the loss of hormones can't be the only thing that causes mental deficits after menopause. Young rhesus monkeys whose ovaries have been removed, depriving them of oestrogen, can still perform well on mental tests, says neuroscientist John Morrison of the Mount Sinai School of Medicine in New York City. "So they could counter the absence of oestradiol, whereas the aged animals could not." Older monkeys can, however, regain robust mental function when given oestradiol supplements.

Oestradiol apparently enhances memory skills by boosting the numbers of dendritic spines, which are sites on neurons where synapses, or connections, form. Morrison's research shows that oestradiol increases the density of synapses in the prefrontal cortex — the region of the brain responsible for the highest level of mental function.

To produce those benefits for ageing monkeys, Morrison and collaborators found, timing is doubly important. Besides administering oestradiol to monkeys near the time of menopause, researchers found the best results when injecting it at 21-day intervals, which mimicked the monkeys' natural hormonal cycle.

On a whim

Timing is clearly at the heart of the debate over HRT in women as well. The large clinical trial — called WHIMS, for the memory-study part of the Women's Health Initiative — documented problems that stemmed from the commonly prescribed combination of oestrogen and progestogen. (Oestrogen alone is typically prescribed only for women who have had their uterus removed, as oestrogen alone raises the risk of uterine cancer.) An analysis of more than 4,000 women found that the rate of dementia was approximately doubled in elderly women on combination HRT compared with those on a placebo.

That surprising result led researchers to ponder whether HRT could be beneficial if administered to younger women, around the time of menopause. And studies examining oestrogen's actions within cells seem to support that suggestion.

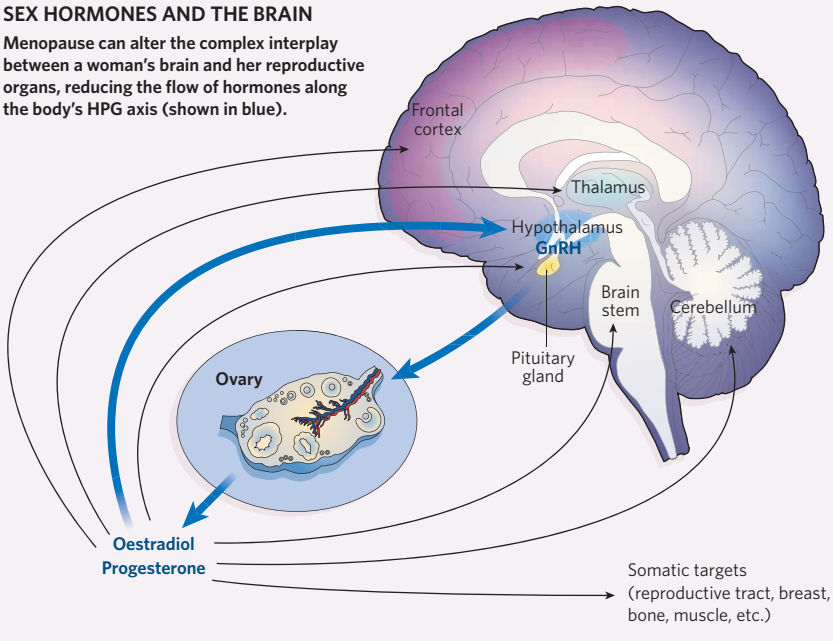
Young, healthy cells are protected by oestrogen, says molecular neuropharmacologist Roberta Brinton of the University of Southern California in Los Angeles. But in cells already affected by disease (say, in the early stages of Alzheimer's), adding oestrogen accelerates the damage, her studies show[5,6]. That could explain why the earlier studies of younger women suggested that HRT helped the brain.

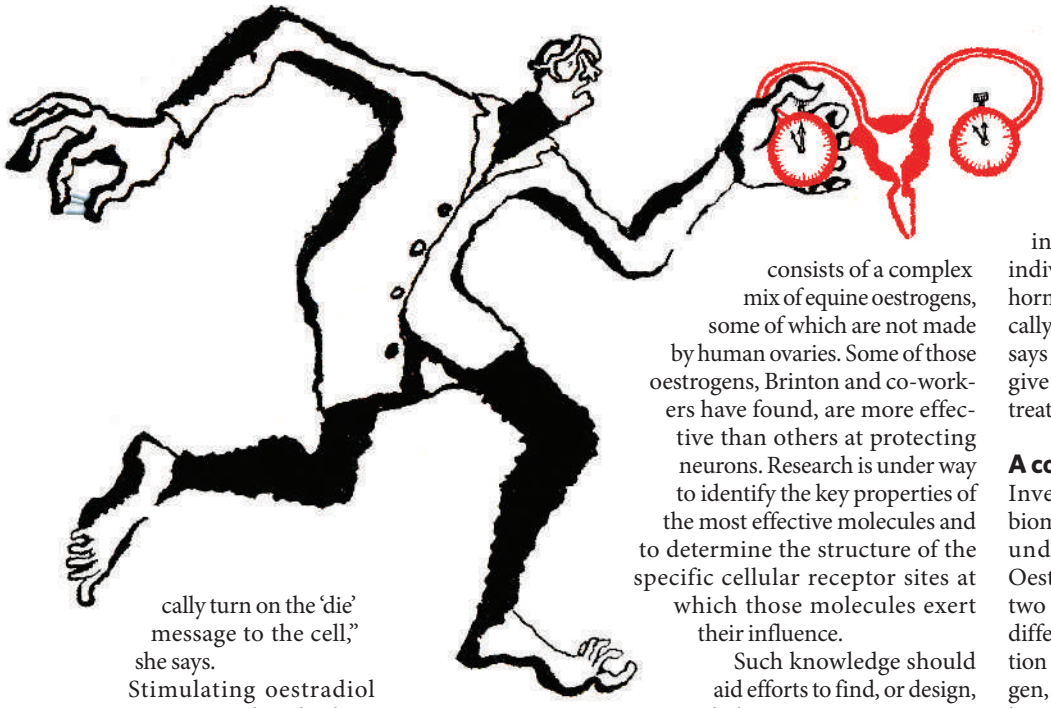
Within a neuron, Brinton says, oestrogen regulates the viability of mitochondria, the cell's energy-production centres. "If you damage mitochondria and you damage enough of them and damage them severely enough, the mitochondria release messengers that basi-

SOURCE: *J. NEUROSCI.*

SEX HORMONES AND THE BRAIN

Menopause can alter the complex interplay between a woman's brain and her reproductive organs, reducing the flow of hormones along the body's HPG axis (shown in blue).





B. MELLOR

cally turn on the 'die' message to the cell," she says.

Stimulating oestradiol receptors on mitochondrial membranes triggers a cascade of reactions that causes calcium to flood into a neuron. For a healthy neuron, that's not a problem, and in fact, the calcium influx initiates other chemical cascades that prevent mitochondria from signalling the cell to die. Older neurons, though, are not able to regulate calcium levels effectively, and therefore suffer even further if given oestrogen.

Defensive driving

"If you take a healthy cell and expose it to oestrogen, you create what I call a proactive defence survival state," Brinton says. Experiments show that such 'protected' cells can resist attacks from toxic substances such as free radicals or amyloid. "These cells have about a 25–50% survival advantage," she says.

On the other hand, for a damaged cell, oestrogen turns from a friend into an enemy. When experimenters expose a cell to toxic insults first, addition of oestrogen later leads to further degeneration[5].

Brinton believes that women in the WHIMS study who developed Alzheimer's disease while taking oestrogen–progestogen supplements were probably in the early stages of Alzheimer's already. Replacement hormones then accelerated the disease.

"That's our hypothesis," she says. "You have to treat these cells in the brain at a time when they are healthy, when they have healthy calcium function, and that allows them to activate these mechanisms that lead to survival." So far, her results are based on studies of cultured neurons, but similar experiments have now started in a transgenic mouse model of Alzheimer's disease.

Further studies in Brinton's lab focus on the precise formulation of HRT, which typically consists of a complex mix of equine oestrogens, some of which are not made by human ovaries. Some of those oestrogens, Brinton and co-workers have found, are more effective than others at protecting neurons. Research is under way to identify the key properties of the most effective molecules and to determine the structure of the specific cellular receptor sites at which those molecules exert their influence.

Such knowledge should aid efforts to find, or design, 'selective oestrogen receptor modulators' for the brain — known as neuroSERMs — that would offer oestrogen's positive neurological benefits without the negative effects of standard HRT. Effective neuroSERMs would penetrate the blood–brain barrier and activate particular oestrogen receptors in the brain, without stimulating receptors responsible for deleterious effects in other tissues, such as in the breast or ovaries. Numerous neuroSERM candidate molecules are being developed and tested[6].

> **"You have to treat these cells in the brain at a time when they are healthy."**
> — Roberta Brinton

Still, for all the advances in mapping out the interplay of brains and hormones, many issues remain unresolved. For instance, more work is needed to determine whether a window of opportunity exists during which people would benefit from HRT, and if so, when it would be best to start, and end, an HRT regimen. "I think nobody really knows the answer," says Schmidt. "It will be important to nail down what we mean by this critical window, and whether in fact it would suffice to take hormone therapy during that time."

Some researchers also think that a similar window of opportunity might exist to evade some of HRT's other possible dangers, particularly heart disease. Oestrogen's effect on atherosclerosis, for instance, will be examined by the new Kronos Early Estrogen Prevention Study, a five-year randomized trial to test oestrogen supplements in women who are within three years of their last menstrual period[7]. The results from the Women's Health Initiative study indicated a raised risk of heart disease for users of HRT, but for much older women.

Ultimately, researchers expect that some women will benefit from HRT whereas others will not. The trick is to identify who belongs to which group. So a major priority of future research will be finding biomolecules that signal whether an individual is likely to benefit (or suffer) from hormone supplements. "Biomarkers are a critically important issue in all of ageing research," says Morrison. "We sorely need biomarkers to give us a reflection of brain changes for any treatment of age-related disorders."

A complicated affair

Investigations aimed at identifying such biomarkers will have to sort out many poorly understood biochemical complications. Oestradiol, for instance, operates on at least two receptors — alpha and beta — found at different sites in the brain. Add in the realization that the brain can produce its own oestrogen, as well as respond to oestrogen from the hormonal system, and ample avenues open up for new subplots in the hormone soap opera.

"You have to take into consideration the complexity of where the oestrogen receptor is," says Brinton. Oestrogen receptors are found not only on mitochondria, but also in a neuron's outer membrane and in its nucleus. And as Gore points out, the body has oestrogen receptors in other tissues too. "There are different tissues and different targets," she says, "and as we age our bodies may express a different assortment of these oestrogen receptors, even within a specific tissue."

And although oestradiol is the most important of the steroidal sex hormones in women, it doesn't work alone. Male sex hormones — the androgens — also circulate, and could affect how much oestradiol ends up acting within the brain, says Gore.

As for men, all the attention given to menopause may have misled them into thinking that they have nothing to worry about with respect to oestrogen and the ageing brain. But hormone and ageing issues may affect them as well. The male brain contains oestrogen receptors too, says Gore. "So there may be similar types of effects in men." ■

Tom Siegfried is a science writer in Los Angeles.

1. Shumaker, S. A. *et al. J. Am. Med. Assoc.* **289,** 2651–2662 (2003)
2. Rapp, S. R. *et al. J. Am. Med. Assoc.* **289,** 2663–2672 (2003).
3. Morrison, J. H., Brinton, A. D., Schmidt, P. J. & Gore, A.C. *J. Neurosci.* **26,** 10332–10348 (2006).
4. Hersh, A. L., Stefanick, M. L. & Stafford, R. S. *J. Am. Med. Assoc.* **291,** 47–53 (2004).
5. Chen, S., Nilsen, J. & Brinton, R. D. *Endocrinology* **147,** 5303–5313 (2006).
6. Zhao, L., O'Neill, K. & Brinton, R. D. *Brain Res. Rev.* **49,** 472–493 (2005).
7. Manson, J. E. *et al. Menopause* **13,** 139–147 (2006).

A switch in time

By 2020 the semiconductor industry wants a memory device that can store a trillion bits of information in an area the size of a postage stamp. As companies race towards this goal, chemists are coming up with an unusual approach. **Philip Ball** reports.

It's a new year. But in the labs of Jim Heath and Fraser Stoddart in California, that year is not 2007, it's 2020. They have leapt into the future with a memory device that potentially matches the needs of the semiconductor industry 13 years from now. The array is no bigger than a single white blood cell, yet it contains 160,000 memory elements, each with an area of just 30 nanometres square — some 40 times smaller than those in existing devices.

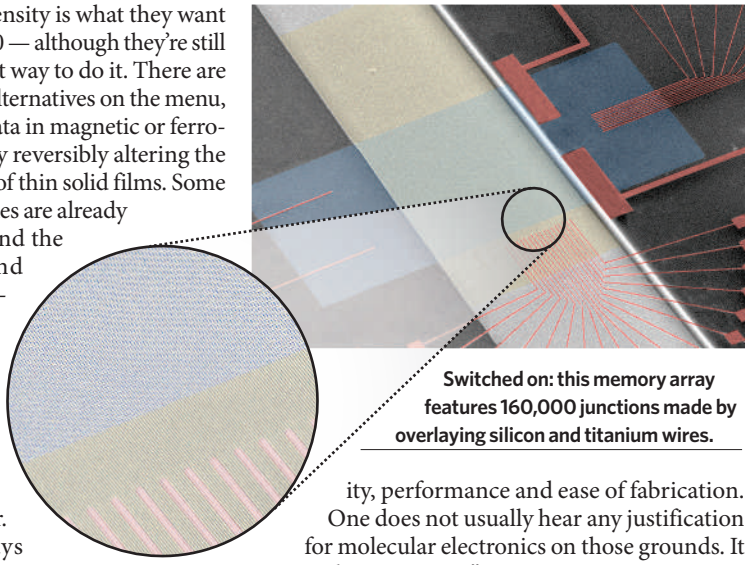
It's not the first time someone has made a prototype ultrasmall memory that is years ahead of its time. But what distinguishes the device made by Heath, Stoddart and their colleagues is that it stores the zeroes and ones of binary information in the switchable states of organic molecules. The system, described in detail on page 414 (ref. 1), brings the idea of molecular memories a step closer to reality.

Although it may have a very high 'bit density' (the number of memory elements per square centimetre), this super-memory isn't going to appear in a laptop any time soon. The researchers are frank about its current shortcomings, not least of which is that the memory cells stop working after being switched just ten times. "It wouldn't surprise me if we really did have to wait until 2020 to see molecular devices with this bit density actually being used," admits Heath, who is based at the California Institute of Technology in Pasadena.

Computer memories have been reaching higher bit densities for decades — but it's becoming ever tougher to keep up with the industry's long-term trends. Today's lithographic techniques for carving silicon into circuit patterns are unlikely to deliver the 2020 target of memory cells just 30 nm or so apart. That's one reason why researchers have begun to think seriously about building memory devices from the bottom up using individual molecules.

"Chemists tend to focus on one aspect of the problem: size."
— Phaedon Avouris

Heath and Stoddart's work is a proof-of-concept, showing that molecular memory cells can be made with a bit density of 10^{11} per cm^2. And it's worth taking seriously, because in terms of miniaturization, the computer companies know that this density is what they want to achieve by 2020 — although they're still unsure of the best way to do it. There are several possible alternatives on the menu, such as storing data in magnetic or ferroelectric cells or by reversibly altering the atomic structure of thin solid films. Some of these approaches are already well advanced, and the route Heath and Stoddart are offering — a curious hybrid of silicon-based microfabrication and organic chemistry — is widely seen as an outside contender.



J. E. HEATH *ET AL.*

Switched on: this memory array features 160,000 junctions made by overlaying silicon and titanium wires.

But Heath says that their silicon circuit has in itself been enough to catch the eye of the semiconductor industry, even though the molecular switches aren't robust, or fast. Stoddart, an organic chemist at the University of California, Los Angeles, who has been exploring switchable molecules since the 1980s, confirms that the interest from industry is real. The work has emerged from the collaborative California NanoSystems Institute, where Stoddart is the director and Heath was the founding director. The institute received four years of initial funding from many IT companies, including $7.8 million from Hewlett-Packard, says Stoddart, and now "Intel has bought into it, to the tune of some $30 million".

But a better memory isn't just about higher bit density. "Chemists tend to focus on one aspect of the problem: size," says Phaedon Avouris of IBM's T. J. Watson Research Center in Yorktown Heights, New York. Although he is also working on molecular devices, made from carbon nanotubes, Avouris stresses that in the end companies such as IBM want something they can make dependably and economically. "The bottom line for devices is always reliability, performance and ease of fabrication. One does not usually hear any justification for molecular electronics on those grounds. It is always just size."

Easy to forget

Today's dynamic random-access memories (DRAMs), the working memory of most electronic devices, leak charge and must be refreshed thousands of times a second, making them power-hungry and draining batteries. What's more, DRAM loses all information when the power is switched off. So power failures can cause loss of data, and a computer's operating system has to be copied afresh from the hard drive during start-up, resulting in long boot-up times.

Ideally a RAM would be non-volatile, meaning that the data don't evaporate the moment power is cut. Flash memory, used in mobile phones and digital cameras, is a form of non-volatile RAM, but it has drawbacks: writing data is very slow, its switching lifetime is limited to around 100,000 cycles, and the data do leak away eventually. That means flash memory isn't a viable option for computers.

The molecular memory made by Heath and Stoddart isn't truly non-volatile yet — the molecules begin to switch states spontaneously after an hour or so. That may be improved by tinkering with the molecular structure to make the two states more stable. But if they hope to replace DRAM or flash they have some way to go to match other non-volatile memories currently being developed.

Magnetic RAM (MRAM) holds data magnetically, so no power is needed to sustain it. The electrical resistance of MRAM cells changes when their magnetic orientation is switched. Last July, the company Freescale Semiconductor in Austin, Texas, released the first commercial MRAM non-volatile memory chip, with a capacity of 4 million bits and a switching time of 35 nanoseconds.

The most mature approach is ferroelectric RAM (FeRAM), where the switching involves altering the polarization state of a ferroelectric material[2]. Commercialization of this technology is fairly advanced: Samsung markets a 64-million-bit FeRAM for low-tech applications such as smart cards. There are already prototype memory arrays with cell spacings of 45 nm — the semiconductor industry's 2010 target for DRAM. Moreover, FeRAMs switch very quickly; commercial devices take just a few nanoseconds. The barriers facing wider adoption of FeRAM are now more economic than technical, says Jim Scott, a specialist in ferroelectrics at the University of Cambridge, UK.

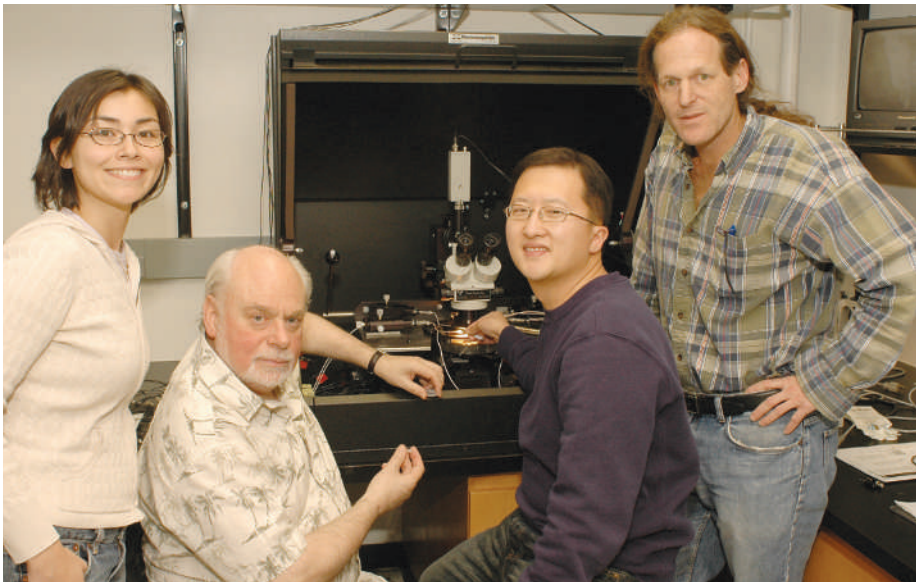
The next phase

Another option for non-volatile RAM is to record data in a 'phase-change material', whose atomic structure can be reversibly altered within a small volume. For example, heating caused by an electric current can partially melt the material, switching it between crystalline and amorphous phases with different electrical conductivity. Commercial 'PRAMs' are already being built by companies such as Samsung and Intel but the bit densities are not much higher than those of today's flash memory.

At the end of last year, IBM researchers unveiled a PRAM device[3] based on a single cell measuring just 3 by 20 nanometres and switching in 2–20 nanoseconds. The team is now working on making large arrays of these elements, something that Spike Narayan, at IBM's Almaden Research Center in San Jose, calls "very feasible". He adds, however, that devices this small can't yet be made with the patterning technologies used for commercial chips.

So Heath and his colleagues face some stiff competition. What do they have to offer? Their array is made up of two sets of tiny wires — one of silicon, the other of titanium — arranged in parallel. Each of these wires is just a few tens of nanometres wide — smaller than the circuit components on today's silicon chips, which are more than 100 nm across. The researchers use a method they devised in 2003 that relies on etching ultrathin layered films rather than lithography to produce the sets of wires. To create the memory array, a set of titanium wires is placed at right angles on top of the silicon wires to form a grid with multiple junctions.

It's at these junctions that the switchable molecules, known as rotaxanes, are anchored. The rotaxanes consist of a linear chain-like molecule threaded through a molecular hoop. The hoop 'docks' at either of two sites along the rotaxane chain, and bulky groups of atoms act as 'stoppers' at the ends. One of these stopper groups is designed to anchor the molecules to silicon, so that they will readily attach themselves to the nanowires. A few hundred rotaxanes at the junction of two wires can be switched by applying voltages to the wires, changing the electrical conductivity of the junction as the molecules become oxidized or reduced and the hoop jumps between the docking sites.

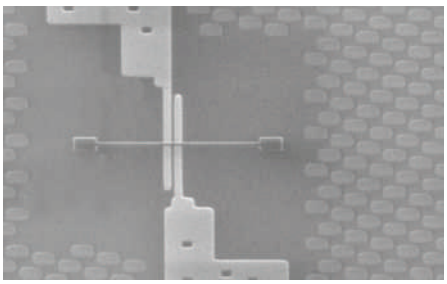


Memory gain: a team including (from left) Bonnie Sheriff, Fraser Stoddart, Jang Wook Choi and Jim Heath has created a high-density memory array that uses organic molecules to store binary data.

Wired up

Heath, Stoddart and their colleagues first demonstrated that these memory cells worked in 2002, using an 8×8 array. Scaling this up to 400×400 nanowires was no easy matter, but their latest memory array has 160,000 junctions, each housing about 100 rotaxane molecules. Wiring up the whole array is challenging, so they have tested a subset of 128 junctions. They found that only half were switchable, and only about half of those gave a sufficiently reliable signal for read-out. In other words, just one in four of the memory elements actually works.

That's not great, but it's not a fatal flaw. Heath and his collaborators have shown that robust memories can be made from defective arrays by using software that finds the 'good' bits and routes around the 'bad' ones[4]. "That isn't so different from magnetic hard-disk memory," says Heath, in which bad sectors are identified so that those bits aren't used. Stoddart suggests the problem is not that the molecules are failing to switch, but that there are limitations to the nanofabrication — especially etching — which he anticipates will improve.



IBM's prototype phase-change RAM is just a few nanometres big and offers very fast switching.

Harry Atwater of the California Institute of Technology, who is working on new types of flash memory, says that "although the rotaxane switch is a triumph for chemical synthetic technique, as a memory element it is orders of magnitude slower than a DRAM cell". Perhaps even more troubling is that ten switching cycles is enough to damage the molecules irreparably. The researchers haven't yet tried to improve the lifetimes — Heath says that automating the fabrication process is their first priority. But he argues that merely making silicon wiring at this density is impressive, and may be useful for miniaturizing standard silicon circuitry beyond memory devices.

For now, high-density PRAM or FeRAM arrays look more likely to supplant DRAMs. Narayan calls the new molecular memory a very fine piece of work, but adds that its application is likely to be different from other arrays. "Molecular memory and PRAM will ultimately cater to very different markets, and are unlikely to compete directly with each other," he says. Rather than being a replacement for conventional DRAM, rotaxane-based arrays might find applications as cheap, disposable memories — which could still be a huge market. For his part, Heath is confident that applications will be found for a device that "only a few years ago people were calling flat-out ridiculous dreams of science fiction". ■

Philip Ball is a consultant editor for *Nature*.

1. Green, J. E. *et al. Nature* **445,** 414–417 (2007).
2. Scott, J. F. & Paz de Araujo, C. A. *Science* **246,** 1400–1405 (1989).
3. Chen, Y. C. *et al.* Paper 30.3 presented at IEEE International Electron Devices Meeting (San Francisco, December 2006).
4. Heath, J. R., Kuekes, P. J., Snider, G. S. & Williams, R. S. *Science* **280,** 1716–1721 (1998).

REF. 3

Journals should set a new standard in transparency

SIR — We applaud your commitment, as expressed in the Editorial "Peer review and fraud" (*Nature* **444,** 971–972; 2006), to raising peer-reviewer awareness about detecting fraud. For studies involving humans, independent research ethics committees (in the United States, institutional review boards) provide the first independent critical scrutiny of research protocols. We recently examined the instructions to authors of 103 medical journals and found that none requires authors to provide to readers (as online supplementary information accompanying the publication) the protocols approved by these committees.

As concern increases about the integrity of published scientific research, we believe that biomedical journals should establish a new standard in human-research transparency. They should require authors to state at submission — and, where judged necessary, in their published articles — that the research has been approved by the relevant ethical committees. All journals publishing research on non-human animals ("Animal experiments under fire for poor design" *Nature* **444,** 981; 2006) should do the same for non-human animal protocols.

Journals should also require authors to provide the full protocols approved by these committees for the editors and peer reviewers, and to allow the journal, if it wishes, to publish these protocols as online supplementary information accompanying publication of the main paper.

Robert P. Dellavalle*†, Kristy Lundahl†, Scott R. Freeman†, Lisa M. Schilling†
*Department of Veterans Affairs Medical Center, 1055 Clermont St, Denver, Colorado 80220, USA
†University of Colorado at Denver and Health Sciences Center, Aurora, Colorado, USA

Wise words from women aren't among top sellers

SIR — Jon Turney's Books and Arts article "Top of the pops" (*Nature* **444,** 819; 2006) begins by asking "What's special about the best popular science books?". From his choice of examples — 11 authors, 13 books — one answer is obvious: they were all written by men.

As a female scientist whose *Brainwashing: The Science of Thought Control* (Oxford Univ. Press, 2004) was one of four books short-listed for the prestigious 2005 *Times Higher Education Supplement* 'Young Academic Author of the Year' award, and as one of two women authors (with Patricia Fara) on the Royal Society's 2005 Aventis Prize long-list of 13, I recognize that Turney was looking at best-selling popular rather than academic books, and that 2006 may have been a leaner year for women authors. And I note that Mary Purton's accompanying "2006 all wrapped up" (*Nature* **444,** 821–822; 2006) includes the novelist Allegra Goodwin and translator Carol Brown Janeway.

Still, I am glad that some organizations, such as the *Times Higher Education Supplement* and the Royal Society, include female science writers among their 'best'.

Kathleen Taylor
Department of Physiology, Anatomy and Genetics, Oxford University, Sherrington Building, Parks Road, Oxford, OX1 3PT, UK

Polluting effects of Brazil's sugar-ethanol industry

SIR — As Brazilian environmental scientists, we believe that the Business Feature "Drink the best and drive the rest"[1] understates the serious environmental and social problems associated with Brazil's sugar-cane ethanol industry. For instance, although you say that soil erosion is a "potentially" damaging side-effect of sugar-cane cultivation, there is abundant scientific evidence already that environmental degradation from soil erosion in sugar-cane fields is widespread[2]. In the state of São Paulo, which is the core of the ethanol industry in Brazil, estimated rates of soil erosion in sugar-cane fields are up to 30 tonnes of soil per hectare per year. Moreover, despite laws to protect the riparian buffers that prevent soil inputs to rivers and streams, only 30% of riparian zones have been preserved in river basins.

The burning of sugar-cane fields before manual harvesting twice a year is another serious environmental problem related to the ethanol industry in Brazil. Although a law passed in 2002 by the state of São Paulo decrees that, by 2006, 30% of the sugar-cane fields with slopes lower than 12% (called mechanizable areas) should not be burned, farmers have been reluctant to replace cheap manual labour with more expensive mechanized harvesting. The deadlines imposed to reduce the burning of sugar-cane fields have been postponed several times, under pressure by sugar-cane farmers. Thus, it is likely that smoke pollution from sugar-cane fields will continue to be a major problem in São Paulo and other Brazilian states for many years, leading to further acidification of the already poor tropical soils[3]. Additionally, high particulate concentrations in the atmosphere from sugar-cane burning have been associated with a growing number of human respiratory diseases in sugar-cane regions[4,5].

Last but not least, although the sugar-cane industry generates jobs in Brazil, working conditions, especially for manual harvesters, are extremely poor and often associated with causes of death. Thus — although we agree that Brazil's ethanol industry is "able to get better" — from an environmental and social standpoint, it is far from being as good as you portray.

We believe that the present ethanol industry and proposals for expansion of ethanol production in Brazil and worldwide should be carefully evaluated, to avoid environmental and social problems far outweighing long-term economic gains.

Luiz Antonio Martinelli*, Solange Filoso†
*Centro de Energia Nuclear na Agricultura, Avenida Centenário 303, 13416-000, Piracicaba, SP, Brazil
†Department of Entomology, University of Maryland, 4112 Plant Sciences Building, College Park, Maryland, 20742-4415, USA

1. *Nature* **444,** 670–672 (2006).
2. Sparovek, G. & Schnug, E. *Soil Sci. Soc. Am. J.* **65,** 1479–1486 (2001).
3. Krusche, A. V. *et al. Environ. Pollut.* **121,** 389–399 (2003).
4. Arbex, M. A. *et al. J. Air Waste Manag. Assoc.* **50,** 1745–1749 (2000).
5. Cançado, J. E. D. *et al. Environ. Health Persp.* **114,** 725–729 (2006).

Magenta and yellow in images is not a bright idea

SIR — Chris Miall, in Correspondence ("Readers see red over low-impact graphics" *Nature* **455,** 147; 2007), points out that a small but significant proportion of the population have difficulty distinguishing colours in red/green images. For this reason, several journals now require images to be published in other colour pairs, the most common being magenta and yellow.

However, when using colours to evaluate protein co-localization, for example, as is now common in confocal microscopy data, some of the alternative colour schemes just do not work. Magenta and yellow in overlay produce 'almost white' — virtually indistinguishable from the yellow in tiny images.

Red and green, the standard colour pair, produce yellow when overlaid, and this is very easy to interpret. I suggest that journals continue to publish these images in red/green, but that they make alternatively coloured images available online as supplementary information for readers who have impaired colour vision.

John Runions
School of Life Sciences, Oxford Brookes University, Gipsy Lane, Oxford OX3 0BP, UK

Contributions to Correspondence may be submitted to corres@nature.com. They should be no longer than 500 words, and ideally shorter. Published contributions are edited.

BOOKS & ARTS

Dark days at the White House

Has the George W. Bush administration manipulated science for political ends?

Undermining Science: Suppression and Distortion in the Bush Administration
by Seth Shulman
University of California Press: 2007.
202 pp. $24.95, £15.95

John Horgan

Two years ago the journalist Ron Suskind offered a disturbing insight into the presidency of George W. Bush. In an article in *The New York Times Magazine* on 17 October 2004, Suskind quoted a senior White House adviser mocking journalists and others in the "reality-based community" who believe that "solutions emerge from your judicious study of discernible reality". The adviser added: "That's not the way the world really works anymore. We're an empire now, and when we act, we create our own reality."

I kept thinking of this statement as I read *Undermining Science*, in which journalist Seth Shulman exposes the Bush administration's attitude towards the "reality-based community" of scientists. Shulman's book expands on a report he compiled for the Union of Concerned Scientists (UCS), a non-profit US watchdog group. Issued in February 2004 during Bush's re-election campaign, the UCS report, entitled 'Scientific Integrity in Policymaking', charged that the Bush administration "is, to an unprecedented degree, distorting and manipulating the science meant to assist the formation and implementation of policy" on a wide range of issues, including climate change, air and water pollution, wildlife protection, sex education, reproductive health, drug abuse, AIDS, workplace safety and missile defence.

The UCS also released a statement in which 62 prominent scientists accused the Bush administration of bending scientific facts to fit its political agenda. Some of the signatories, such as the physicist Richard Garwin, a designer of the hydrogen bomb, had advised Republican as well as Democratic administrations. The report and statement — now signed by more than 10,000 scientists — triggered widespread commentary. In an editorial on 26 April 2004 entitled 'Bush-league Lysenkoism', the normally staid *Scientific American* declared: "It is increasingly impossible to ignore that this White House disdains research that inconveniences it."

Shulman wrote *Undermining Science* "to underscore the issues at stake in more explicit and personal language". Actually, his rhetoric is mild compared with that of Chris Mooney, whose book *The Republican War on Science* (Basic Books, 2005) lambasts not just Bush officials but Republicans in general for their attitude towards science. Whereas Mooney delivers scathing profiles of such key figures as Bush's science adviser John Marburger, Shulman sticks more or less to recounting the facts. That makes *Undermining Science* a valuable complement to *The Republican War on Science*. Shulman's book serves as a concise, straightforward case history of the politicization of science, ideal for courses on the history, philosophy, sociology and ethics of science.

M. HUMPHREY/AP

The administration of George W. Bush has been widely attacked over its approach to science.

Shulman acknowledges that other administrations have introduced political calculations into scientific and medical matters. As the journalist Stephen Hall documents in his book *Merchants of Immortality* (Houghton Mifflin, 2003), the Bill Clinton administration impeded federal research on embryonic stem cells for fear of offending religious voters. Clinton also fired his surgeon general, Jocelyn Elders, in 1994 after she suggested that sex education should include information about masturbation.

Moreover, scientists are not always paragons of objectivity. Shulman quotes Harvard biologist Richard Lewontin: "Why should we trust scientists, who, after all, have their own political and economic agendas?" But dwelling on run-of-the-mill scientific bias in the context of the Bush administration's behaviour, Shulman argues, "is rather like conducting an argument about the extent to which pilots normally deviate from their flight plan while riding in an airplane that has just been hijacked". Indeed, Lewis Branscomb, head of the National Bureau of Standards under the Republican President Richard Nixon, is quoted in *The Christian Science Monitor* on 6 January 2004 as saying: "I don't think we've had this kind of cynicism with respect to objective scientific advice since I've been watching government, which is quite a long time."

A few key episodes recounted by Shulman give a sense of the brazenness, if not the enormous scale, of the Bush administration's tactics. In 2004, the White House dismissed the eminent biologist Elizabeth Blackburn, an outspoken proponent of research on embryonic stem cells, from the President's Council on Bioethics. Blackburn was replaced by Diana Schaub, a political scientist who described research that led to the destruction of embryonic stem cells as "evil".

The White House has also staffed its scientific agencies with people who openly question the agencies' missions and punish employees for doing their jobs. Craig Manson, assistant

secretary for fish and wildlife and parks, once stated: "If we are saying that the loss of species in and of itself is inherently bad — I don't think we know enough about how the world works to say that." In 2004, on the day after Bush's re-election, the Fish and Wildlife Service, which Manson oversees, fired the biologist Andy Eller, who had charged that the service was not fulfilling its mandate of protecting the Florida panther, an endangered species. The agency was forced to re-hire Eller after a court ruled in his favour.

Some of the Bush administration's actions have been almost comically incompetent. Last January, for example, George Deutsch, a public-affairs officer at NASA, tried to prevent the space agency's James Hansen from speaking to the press about the dangers of global warming. Andrew Revkin of *The New York Times* quickly exposed the attempt to censor Hansen, and it was soon revealed that Deutsch, contrary to what it said on his CV, had never graduated from university. Deutsch resigned and NASA administrator Michael Griffin declared: "It is not the job of public-affairs officers to alter, filter or adjust engineering or scientific material produced by NASA's technical staff."

But the damage caused by the Bush administration's contempt for scientific facts is no laughing matter. In the two years since Bush was re-elected, "reality" — especially the reality of Iraq, which, Shulman points out, the United States invaded on the basis of erroneous technical claims — has humbled his administration. In November, as the death toll in Iraq surged, voters handed the reins of power in Congress to the Democrats. In December, the Iraq Study Group, headed by the Bush family friend James Baker, issued a scathing critique of the US occupation of Iraq. Bush's approval rating has sunk to one of the lowest levels ever recorded. The declaration of that Bush official — "We create our own reality" — has now taken on a tragic irony. ■

John Horgan is director of the Center for Science Writings at the Stevens Institute of Technology, Hoboken, New Jersey. His most recent book is *Rational Mysticism*.

The Universe's quantum monkeys

Programming the Universe: A Quantum Computer Scientist Takes On the Cosmos
by Seth Lloyd
Alfred A. Knopf/Jonathan Cape: 2006.
240 pp. $25.95/£18.99

Artur Ekert

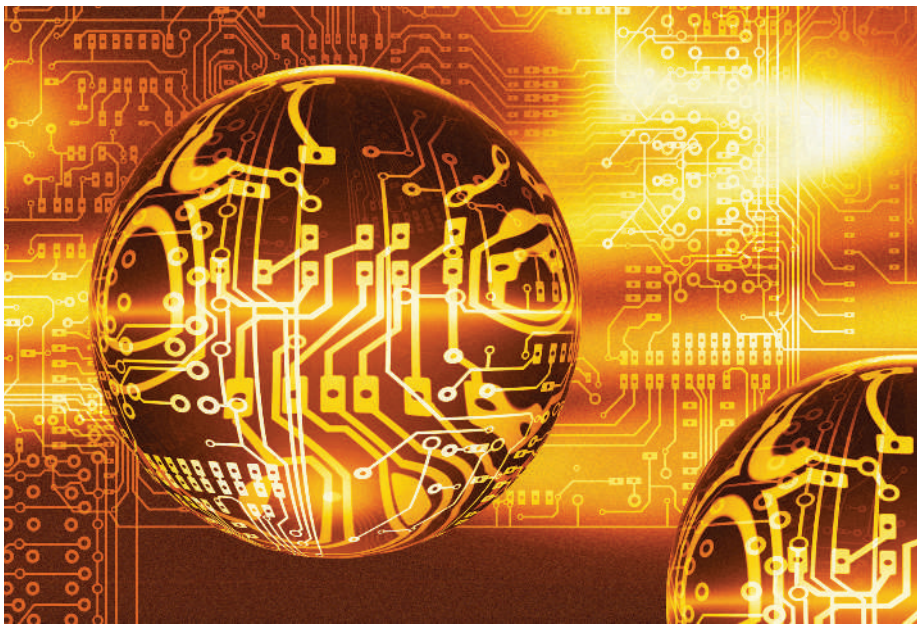
A little less than 14 billion years ago, a huge explosion gave birth to the Universe, and once it sprang into existence, the Universe began computing. The positions, velocities and internal states of every elementary particle, every atom and molecule, indeed every single physical entity register bits of information. Those bits are continually altered by physical interactions that act like sequences of logic gates — given a sufficient supply of bits and enough time, they can compute just about anything that is computable. Thus, the Universe is a computer. It is not a metaphor, it really is. More than that, the fundamental laws of physics that govern any interaction are quantum; hence, the Universe is a huge quantum computer that computes its own behaviour. It started in a very simple state initially, but in time, as the number of computational steps increased, the computing quantum Universe spun out more complex patterns, including galaxies, stars and planets, and then life, humans, you and me, and Seth Lloyd and his book *Programming the Universe*.

Like many other good stories of this type, Lloyd's book will puzzle and even irritate as much as it persuades. Lloyd writes in a lively style, weaving jokes and personal anecdotes into more technical narrative. He shares his views on cosmology, computation, quantum physics, complexity, sex, life, the Universe and all that, and he does it well. Despite this proliferation of topics, the main message stands out and is reiterated several times — the Universe is a quantum computer programmed by quantum fluctuations, and the computational capability of the Universe explains how complex systems can arise from fundamentally simple physical laws.

Lloyd tells the story of the evolving Universe in terms of interplay between energy and information. In the conventional history of the origin and the evolution of the Universe, the story usually told by cosmologists and astronomers, energy plays the central role. First there was a singularity and there was no past for it to emerge from. Then expansion. As the Universe expanded, it cooled down and various forms of matter condensed out because the disruptive thermal energy gradually dropped below the binding energies that hold constituent parts of protons, nuclei and atoms together. Tiny quantum fluctuations made some regions of the Universe slightly denser, and gravity amplified this effect, which resulted in gas clouds, stars and galaxies. Stars exploding to supernovae produced heavier elements, then our Sun and Solar System formed, and about 4 billion years ago, life emerged on Earth. But this story leaves many questions unanswered. How did life arise? Why is the Universe so complex? Could such complexities have arisen from total randomness?

Now, enter computer science. Algorithmic information theory shows that there are short, random-looking programs that can cause a computer to produce complex-looking outputs. Lloyd illustrates this with a popular story attributed to the French mathematician Émile Borel. Imagine a bunch of monkeys typing randomly into typewriters. Given enough time, it is certainly possible that one of these monkeys will type the first million digits of π or the first act of *Hamlet*. Possible, but very unlikely. Now, take the typewriters away and give the monkeys computers that recognize any random inputs not as text but as a computer program. When the computers try to execute random programs, most of the time they will crash or generate garbage, but every now and then just a few lines of random code typed by monkeys will give interesting outputs — for example, the successive digits of π, or intricate fractals. Or perhaps much more interesting patterns if the computer is the Universe itself.

This vision of a computational Universe is



M. KULYK/SPL

Get with the program: the Universe, it turns out, is actually a giant quantum computer.

not new: it was proposed in the 1960s by Konrad Zuse and Ed Fredkin, and revived more recently by Stephen Wolfram. However, unlike his predecessors, Lloyd stresses the quantum nature of computation. This distinction is important because, to the best of our knowledge, it seems impossible to simulate the evolution of a quantum system in an efficient way on a classical computer.

A classical computer simulation of quantum evolution typically involves an exponential slowdown in time. This is because the amount of classical information needed to describe the evolving quantum state is exponentially larger than that needed to describe the corresponding classical system with similar accuracy. However, instead of viewing this intractability as an obstacle, today we regard it as an opportunity — if that much computation is needed to work out what will happen in a quantum multi-particle interference experiment, then the very act of setting up such an experiment and measuring the outcome is equivalent to performing a complex computation. Since Richard Feynman and David Deutsch pointed out this opportunity in the 1980s, the hunt has been on for interesting things for quantum computers to do, and at the same time, for the scientific and technological advances that could allow us to build quantum computers. The field is flourishing, and Lloyd provides a good popular introduction to the subject. However, he does not stop at the level of building quantum computers, he takes on the biggest quantum computer there is — the Universe.

The Universe is a quantum computer, and quantum mechanics supplies the Universe with 'monkeys' in the form of ubiquitous random quantum fluctuations — the same fluctuations that provided the seeds of galaxy formation and of all that followed. The Universe has pockets of complex behaviour because, Lloyd claims, the monkeys have been working very hard. He estimates that the visible Universe, programmed by quantum fluctuations, has performed about 10^{122} operations on 10^{92} quantum bits. No wonder we are here!

I think this is a delightful book, but some parts are patchy and many details are brushed under the carpet. For example, anyone trying to work out numerical estimates of the physical limits to computation or the computational capacity of the Universe is much better off consulting Lloyd's original paper on the subject (see *Nature* **406,** 1047–1054; 2000). It is clear that Lloyd has forsaken accuracy for snappiness in several places, but then this is a popular exposition.

Seth Lloyd is a good storyteller, but is the story convincing? Well, I was convinced, but when I tried a nice line from the book — "programmed by quanta, physics gave rise to chemistry and then to life, programmed by mutation and recombination, life gave rise to Shakespeare, programmed by experience and imagination, Shakespeare gave rise to *Hamlet*" — on a colleague of mine, an English literature fellow, he only shook his head in disbelief and walked away. ■

Artur Ekert is at the Mathematical Institute, University of Oxford, UK, and the National University of Singapore.

Cover story

Skin: A Natural History
by Nina G. Jablonski
University of California Press: 2006.
290 pp. $24.95, £15.95

John Galloway

Biology is a historical science. Ask a 'why?' question about biology, as Nina Jablonski keeps doing in her book *Skin*, and you invite an evolutionary answer. She also tells us everything we might want to know about skin; perhaps more than some people want to know. She then goes on to take informed guesses as to why skin is the way it is and, by implication, why it is not like something else. Skin's appearance, its form and function, questions of how and why it works, and sometimes doesn't, have been thrashed out over a billion or so years at — to borrow her words — the "negotiating table of evolution".

For Alexander Pope, the "proper study of mankind" may have been "man", but Jablonski, as befits a modern biologist, thinks otherwise. Understanding starts, and possibly finishes, with comparisons, between humans and our biological relatives and neighbours, both near and not-so-near. We may share virtually all our genes with chimpanzees, but those we don't share are responsible for a lot of differences, reproductive, linguistic and cognitive. Skin genes, for example, which are responsible among other things for colour, body hair and the number of sweat glands, may well explain why chimps are still confined to African jungles, whereas we, their closest relatives, have already been to the Moon.

Skin is not just about biology, but also the way we live. Our skin is the visible, immediate personal territory where biology most obviously gives way to culture. Jablonski quotes Franz Kafka, who had the right idea, viewing the skin as "not only a garment but also a straitjacket and fate". People go to a lot of trouble and expense to alter their appearance and change their fate. From war-paint and cosmetics to tanning, bleaching, tattooing, ritual scarring, botox, body piercing and 'nipping and tucking', there is someone making money out of it. And it does not necessarily have to stop just because someone is dead, as some enterprising Ancient Egyptian undertaker realized.

The skin's characteristics have been thrashed out at "the negotiating table of evolution".

Some forms of personal make-over and disguise teach a salutory lesson: that culture comes at a biological price, paid from the genetic legacy bequeathed you by evolution. You are, let's say for the sake of argument, a fair-skinned northern European. But it has become the thing to show off a nice tan (Jablonski fingers fashionista Coco Chanel as the perpetrator of this particular vanity), and that means lying about without clothes in hot sun in latitudes rather nearer the Equator. The trouble is, the reason you are fair is a good historical one, indeed a matter of life and death for your ancestors in the Europe of 50,000 or so years ago. And that fact has implications for modern day Sun-worshippers, some of whom discover that mortality still starts with the skin.

At the core of Jablonski's theme is the skin's ability to multi-task: it protects, controls temperature, senses the world around you, and shows people how you really feel, as opposed to what you choose to tell them. But skin is also a chemical factory, fuelled in part by solar radiation. It manufactures vitamin D, without which you can neither extract calcium from your diet nor incorporate it in your bones — posing something of a challenge to survival. Here's an evolutionary conundrum. Ultraviolet light, which damages DNA directly and also destroys the folic acid essential for its synthesis, is, ironically, the energy source needed to make vitamin D. In equatorial Africa, our ancestral home, evolution engineered a nice compromise that allowed humans to leave the sheltering forest canopy and begin global colonization. Melanins that absorb ultraviolet

afforded enough protection for the DNA, but left scope for the necessary production of vitamin D. When our dark-skinned ancestors started to migrate to the rest of the world, they first colonized regions that also had the strong sunlight to which their skin was already well adapted.

But dark skin was not adapted to the lower intensity of sunlight in northern Europe. It was simply over-protected, leading to problems producing vitamin D. Jablonski argues that Europe could only have been colonized in the wake of a genetic mutation that altered both the nature and the distribution of melanins in skin, producing fair skin with a tendency for freckles. The European climate selected for a gene that might well have been lethal back in equatorial Africa. Evolutionary negotiation achieved a new compromise.

These historical events have reverberated down the years, from biological prehistory into human documented history. On the one hand there have been rocketing frequencies of skin cancers in fair-skinned people exposed to too much sun; on the other, rickets became a problem in both the white-skinned populations of sun-deprived, smoke-polluted industrial Britain, and the later, darker-skinned immigrants to a postindustrial but still relatively unsunny northern Europe. These issues hint at the medical truth that any deep understanding of our ideas of 'wellness' and 'illness' is only likely to come from the central concepts of evolutionary theory: reproductive fitness and adaptation. It is amazing that medicine does not make much more use of evolutionary ideas. It is surely a sea change that is long overdue. ■

John Galloway is at the Eastman Dental Hospital, 256 Gray's Inn Road, London WC1X 8LD, UK.

SCIENCE IN CULTURE

Chart toppers

An exhibition explores the diverse ways of putting data on the map.

Martin Kemp

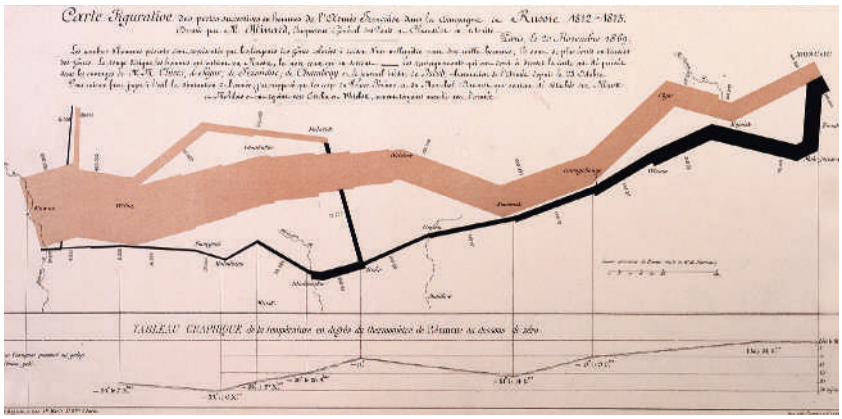
It was often said that geography was about maps, and history about chaps. But there are virtually no sets of data — about chaps or anything else — that cannot be mapped, although sometimes a visually appealing map can hide as much as it reveals. This is the message of the exhibition 'Places and Spaces: Mapping Science' (http://scimaps.org/exhibit/nyscience/), which can be seen at the New York Hall of Science until 25 February, after which it will tour the United States, Japan and Europe.

Curated by Katy Börner and Julie Smith of Indiana University, the exhibition contains a judicious selection of early and modern maps that lay out in various ways how historic cultures have charted the configuration of the Earth and the heavens. There are also specialist maps showing, for example, the distribution of telegraphic linkages, exports and poverty. The charts of scientific papers and patents by region are especially revealing, the latter showing the globe grotesquely morphed by 'fat' zones of high innovative activity.

Relationships between 'chaps' are also mapped, with networks of scientists and citations to the fore. Scientific disciplines are laid out according to their apparent relationships, and historical episodes are charted over time and space, most notably the discovery of DNA. Internet activity is plotted both architecturally and dynamically.

As the entities in the charts become more conceptual, the challenges become more complex and ultimately intractable. One set of juxtapositions in two dimensions necessarily precludes others that may be equally valid. An attempt to devise a two-dimensional distribution of geographical terms ends up by separating space in its political dimension from urban planning, which cannot be right.

The non-topographical mapping seems distinctively modern. But one of the most effective statistical maps is also the earliest in the exhibition. In 1869, the retired French engineer Charles Joseph Minard produced a remarkable map of Napoleon's catastrophic invasion of Russia. It has hardly been surpassed for visual efficacy, and was acclaimed by the pioneer of 'cinematographic' photography, Etienne-Jules Marey, for its "brutal eloquence".



E. R. TUFTE, GRAPHICS PRESS

Decline and fall: Charles Joseph Minard mapped French losses in Napoleon's invasion of Russia.

The basis of Minard's map (shown here) is a straightforward chart of the territory traversed by the French army in the winter of 1812–13, from Kowno (Kaunas) on the left to Moscow on the right. The width of the tinted strip represents the number of soldiers in the French army as it progresses eastwards. The plot doesn't thicken — quite the reverse in fact, beginning with a rearguard of 33,000 men branching off to remain at Polotsk. A series of abrupt falls then brings the company of 422,000 down to 100,000 by the time they reach Moscow.

The black line represents the retreating body of soldiers after they had sacked the Russian capital. The diminishing band is temporarily boosted from 20,000 to 50,000 as the rearguard rejoins, but at the River Berezina a Russian attack triggers a shambles. A stark black step graphically records the extent of the disaster.

Even more remarkably, Minard also charts one of the causal factors, the severe drop of temperature, on the army's return march during November and early December. By 6 December it had reached –30 °C and the army had been scythed down to a tiny rump of just 12,000 men.

Minard's map is both a vivid graphic and a tool for analysis, and forms the basis for further questions, such as the onset of infections, food deprivation, failure of equipment, and shortage of ammunition. Minard's map lays down a series of challenges in lucidity, functionality and potentiality that few since have fully met.

All this was accomplished by Minard after the age of 70. His last act was to publish the Napoleonic map in juxtaposition to one showing the Roman emperor Hadrian's disastrous losses on his return trip across the Alps from his northern expedition. It was designed to demonstrate the enduring human cost of war.

Martin Kemp is professor of the history of art at the University of Oxford, Oxford OX1 1PT, UK. His new book, *Seen | Unseen*, is published by Oxford University Press.

Biology's next revolution

The emerging picture of microbes as gene-swapping collectives demands a revision of such concepts as organism, species and evolution itself.

Nigel Goldenfeld and Carl Woese

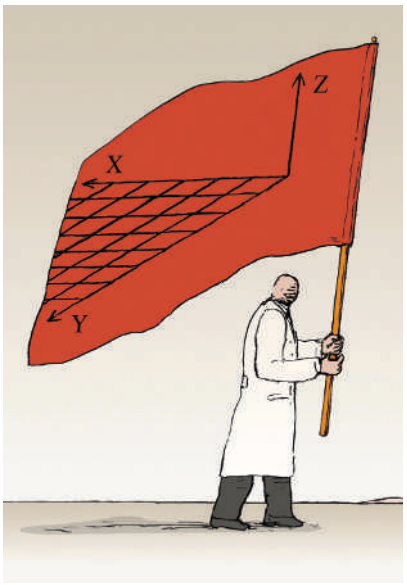
One of the most fundamental patterns of scientific discovery is the revolution in thought that accompanies a new body of data. Satellite-based astronomy has, during the past decade, overthrown our most cherished ideas of cosmology, especially those relating to the size, dynamics and composition of the Universe.

Similarly, the convergence of fresh theoretical ideas in evolution and the coming avalanche of genomic data will profoundly alter our understanding of the biosphere — and is likely to lead to revision of concepts such as species, organism and evolution. Here we explain why we foresee such a dramatic transformation, and why we believe the molecular reductionism that dominated twentieth-century biology will be superseded by an interdisciplinary approach that embraces collective phenomena.

The place to start is horizontal gene transfer (HGT), the non-genealogical transfer of genetic material from one organism to another — such as from one bacterium to another or from viruses to bacteria. Among microbes, HGT is pervasive and powerful — for example, in accelerating the spread of antibiotic resistance. Owing to HGT, it is not a good approximation to regard microbes as organisms dominated by individual characteristics. In fact, their communications by genetic or quorum-sensing channels indicate that microbial behaviour must be understood as predominantly cooperative.

In the wild, microbes form communities, invade biochemical niches and partake in biogeochemical cycles. The available studies strongly indicate that microbes absorb and discard genes as needed, in response to their environment. Rather than discrete genomes, we see a continuum of genomic possibilities, which casts doubt on the validity of the concept of a 'species' when extended into the microbial realm. The uselessness of the species concept is inherent in the recent forays into metagenomics — the study of genomes recovered from natural samples as opposed to clonal cultures. For example, studies of the spatial distribution of rhodopsin genes in marine microbes suggest such genes are 'cosmopolitan', wandering among bacteria (or archaea) as environmental pressures dictate.

Equally exciting is the realization that viruses have a fundamental role in the biosphere, in both immediate and long-term evolutionary senses. Recent work suggests that viruses are an important repository and



KAPUSTA

memory of a community's genetic information, contributing to the system's evolutionary dynamics and stability. This is hinted at, for example, by prophage induction, in which viruses latent in cells can become activated by environmental influences. The ensuing destruction of the cell and viral replication is a potent mechanism for the dispersal of host and viral genes.

It is becoming clear that microorganisms have a remarkable ability to reconstruct their genomes in the face of dire environmental stresses, and that in some cases their collective interactions with viruses may be crucial to this. In such a situation, how valid is the very concept of an organism in isolation? It seems that there is a continuity of energy flux and informational transfer from the genome up through cells, community, virosphere and environment. We would go so far as to suggest that a defining characteristic of life is the strong dependency on flux from the environment — be it of energy, chemicals, metabolites or genes.

Nowhere are the implications of collective phenomena, mediated by HGT, so pervasive and important as in evolution. A computer scientist might term the cell's translational apparatus (used to convert genetic information to proteins) an 'operating system', by which all innovation is communicated and realized. The fundamental role of translation, represented in particular by the genetic code, is shown by the clearly documented optimization of the code. Its special role in any form of life leads to the striking prediction that early life evolved in a lamarckian way, with vertical descent marginalized by the more powerful early forms of HGT.

Refinement through the horizontal sharing of genetic innovations would have triggered an explosion of genetic novelty, until the level of complexity required a transition to the current era of vertical evolution. Thus, we regard as regrettable the conventional concatenation of Darwin's name with evolution, because other modalities must also be considered.

This is an extraordinary time for biology, because the perspective we have indicated places biology within a context that must necessarily engage other disciplines more strongly aware of the importance of collective phenomena. Questions suggested by the generic energy, information and gene flows to which we have alluded will probably require resolution in the spirit of statistical mechanics and dynamical systems theory. In time, the current approach of post-hoc modelling will be replaced by interplay between quantitative prediction and experimental test, nowadays more characteristic of the physical sciences.

Sometimes, language expresses ignorance rather than knowledge, as in the case of the word 'prokaryote', now superseded by the terms archaea and bacteria. We foresee that in biology, new concepts will require a new language, grounded in mathematics and the discoveries emerging from the data we have highlighted. During an earlier revolution, Antoine Lavoisier observed that scientific progress, like evolution, must overcome a challenge of communication: "We cannot improve the language of any science without at the same time improving the science itself; neither can we, on the other hand, improve a science without improving the language or nomenclature which belongs to it." Biology is about to meet this challenge. ■

Nigel Goldenfeld is in the Department of Physics and Institute for Genomic Biology, University of Illinois at Urbana-Champaign, 1110 West Green Street, Urbana, Illinois 61801, USA. Carl Woese is in the Department of Microbiology and Institute for Genomic Biology, 601 South Goodwin Avenue, Urbana, Illinois 61801, USA.

FURTHER READING
Frigaard, N., Martinez, A., Mincer, T. & DeLong, E. *Nature* **439,** 847–850 (2006).
Sullivan, M. *et al. PLoS Biol.* **4,** e234 (2006).
Pedulla, M. *et al. Cell* **113,** 171–182 (2003).
Vetsigian, K., Woese, C. & Goldenfeld, N. *Proc. Natl Acad. Sci. USA* **103,** 10696–10701 (2006).

For other essays in this series, see http://nature.com/nature/focus/arts/connections/index.html

NEWS & VIEWS

MATHEMATICAL PHYSICS

On the right scent

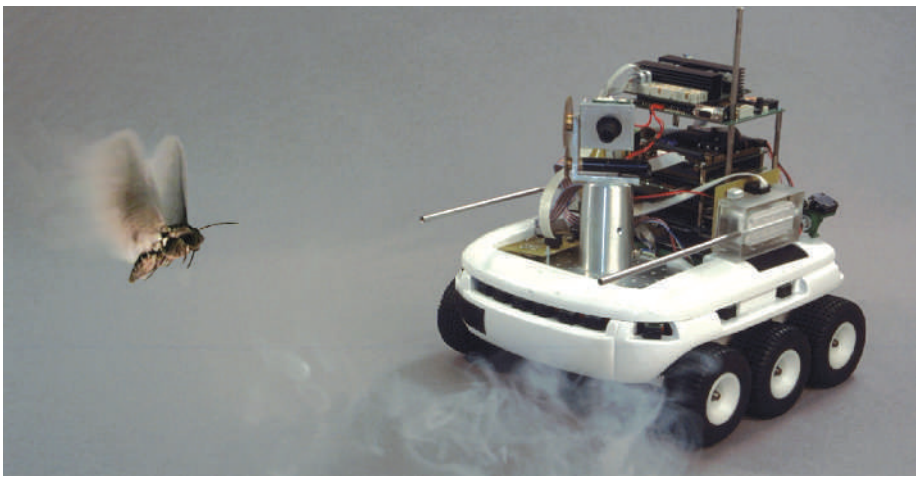
Dominique Martinez

Searching for the source of a smell is hampered by the absence of pervasive local cues that point the searcher in the right direction. A strategy based on maximal information could show the way.

Like many other insects, a female moth releases specific blends of odours — pheromones — to signal her presence to males. The pheromone plume is not a smooth, continuous cloud but consists of intermittent, wind-blown patches of odour separated by wide voids. The probability of encountering one of these patches decays exponentially with distance from their source. Under these circumstances, finding the pheromone source cannot be accomplished simply by 'chemotaxis' — climbing a chemical concentration gradient[1]. On page 406 of this issue[2], Vergassola, Villermaux and Shraiman bring a breath of fresh air to the olfactory search problem: given scanty information, how do moths successfully locate their mates over distances of hundreds of metres?

The problem of searching for odour sources with scarce information is common not only to many air- and water-borne animals[3], but also to olfactory robots designed to search for chemical leaks, drugs and explosives[4] (Fig. 1). Because of the random nature of the odour plume, an exact model of the environment is not available in these cases. Vergassola and colleagues' search scheme[2] circumvents this problem by relying on a probability map of source location that is estimated from the available sensory information — the history of detection and non-detection events.

In an olfactory search, the searcher is initially far from the source and senses small patches of odour. The estimated spatial probability distribution of the source location is consequently flat and broad, and the entropy of the distribution high, reflecting the searcher's uncertainty about the source location. At this stage of a search, making a beeline for locations of maximal estimated probability is a risky strategy. The searcher should instead explore the local environment and gather information so as to obtain a more reliable estimate of the source probability distribution. As the searcher encounters odour particles and accumulates information along its path, the estimated source distribution becomes sharper, and its entropy decreases. Exploitation can now gradually replace exploration, and the searcher can direct itself towards those locations where the probability of finding the source is greatest.



D. MARTINEZ

Figure 1 | Scent trackers. Inspired by the problem of animal navigation in odour plumes, Vergassola *et al.*[2] propose a search algorithm that might help engineers to design efficient olfactory tracking robots.

Striking the right balance between exploration and exploitation is the key to efficient searching when information is hard to come by.

Vergassola *et al.*[2] show that the expected time taken to complete such a search is indeed determined by the entropy of the source distribution. A reduction in the entropy of the estimated distribution is thus a necessary (but not sufficient) condition for an effective search. The authors propose a search algorithm that they term infotaxis. The idea of this algorithm is to take whatever action — making a move or staying still — that maximizes the expected reduction in entropy of the source probability field, and therefore the rate of information acquisition.

Maximizing information gain has been used previously for exploration tasks carried out by autonomous robots[5]. For olfactory searches, Vergassola *et al.* derive a mathematical expression for the variation in entropy that consists of two terms. The first term is identified as exploitative, because it drives the searcher towards points at which there is a high likelihood of finding the source. The second term is explorative, as it favours motion to regions with lower probabilities of source discovery, but high rewards in terms of information useful for improving the estimation of the source distribution. Infotaxis thus naturally combines exploitation and exploration by taking into account both the direct gain specific to finding the source and the knowledge gain from receiving additional olfactory cues. The authors' simulations with modelled and experimental data from turbulent flows indicate that the infotactic search model is significantly faster than explorative or exploitative searches taken in isolation.

The authors note that simulated infotactic paths share qualitative similarities with trajectories observed in the flights of birds and moths. Flights upwind performed by moths attracted by a sexual pheromone have been described extensively at the behavioural level by neuroethologists. When a male moth detects the pheromone of a female, it does not fly straight ahead towards the source, but steers in a zigzag upwind along the pheromone plume. Whenever the moth cannot identify the pheromone plume, it draws wide loops or crosswind turns without forward movement, a tactic known as casting[6]. This odour-modulated behaviour has in the past inspired simple models of olfactory search[7,8]. The striking feature of the infotaxis model is that the casting and zigzagging steps are not preprogrammed by imposing explicit rules of movement such as 'advance upwind' or 'turn crosswind'. Rather, they both emerge naturally from locally

maximizing information gain. When navigating in turbulent odour plumes, it seems that the rate of information acquisition could have a similar role to that of the concentration gradient in chemotaxis.

Although the simulated infotactic trajectories resemble their biological counterparts, the control mechanisms underlying the similarities in trajectories might well differ. In moth flights, for example, the temporal regularity of the turns, whether expressed in zigzagging upwind or in casting, suggests the existence of an internal oscillatory mechanism, known as self-steered counterturning[9]. Search models based on counterturning produce trajectories similar to those observed for moths in wind-tunnel experiments[7,8], and might also to some extent account for the complex 'Lévy-flight' patterns characteristic of insect flights in field studies[10].

But Vergassola *et al.* did not develop their search algorithm on the basis of control mechanisms specific to moths. They considered the problem of olfactory search as sufficiently universal that maximizing information gain allows any searcher to track a chemical plume efficiently to its source. This hypothesis is plausible, because many animals, including crabs and birds, exhibit zigzagging trajectories very similar to those of moths, even though these are probably subject to completely different control mechanisms[3]. In infotaxis, crosswind casting and zigzagging upwind can be viewed as parts of a behavioural continuum ranging from pure exploration to pure exploitation.

The authors' work[2] is intriguing in several respects, and will certainly foster future research. Perhaps the most direct implication is the potential use of infotaxis in robotic search applications, in part because computationally efficient algorithms that update a source probability map in real time are already available for on-board implementation[11]. ■

Dominique Martinez is at the Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA-CNRS), Campus Scientifique, 54506 Vandoeuvre-Lès-Nancy, France.
e-mail: dominique.martinez@loria.fr

1. Berg, H. C. *Nature* **254,** 389–392 (1975).
2. Vergassola, M., Villermaux, E. & Shraiman, B. I. *Nature* **445,** 406–409 (2007).
3. Vickers, N. J. *Biol. Bull.* **198,** 203–212 (2000).
4. Marques, L. & de Almeida, A. (eds) *Auton. Robots* **20,** 183–287 (2006).
5. Thrun, S., Burgard, W. & Fox, D. *Probabilistic Robotics* (MIT Press, 2005).
6. Kennedy, J. S. & Marsh, D. *Science* **184,** 999–1001 (1974).
7. Belanger, J. H. & Arbas, E. A. *J. Comp. Physiol. A* **183,** 345–360 (1998).
8. Balkovsky, E. & Shraiman, B. I. *Proc. Natl Acad. Sci. USA* **99,** 12589–12593 (2002).
9. Kennedy, J. S. *Physiol. Entomol.* **8,** 109–120 (1983).
10. Reynolds, A. M. *Phys. Rev. E* **72,** 041928 (2005).
11. Pang, S. & Farrell, J. A. *IEEE Trans. Syst. Man Cybern. B* **36,** 1068–1080 (2006).

ATOMIC PHYSICS

The social life of atoms

Maciej Lewenstein

In a trail-blazing experiment 50 years ago, it was observed that photons from far-off stars bunch up. But in fact there's a more general distinction among free, non-interacting particles: bosons bunch, and fermions 'antibunch'.

Counting individual quantum-mechanical objects such as the particles of a complex many-body system — whether photons, electrons, atoms or something else — is an efficient way to learn about the properties of both the system and of the particles being counted. Fifty years ago, Robert Hanbury Brown and Richard Twiss[1] published the results of the paradigmatic experiment of this sort, in which they counted joint detections, in two separate detectors, of photons from distant stars. The two-photon correlations clearly showed that the photons liked to arrive bunched up in groups. Jeltes *et al.*, whose results appear on page 402 of this issue[2], use less well-travelled particles for their investigations — ultracold helium atoms. But they are able, for the first time in the same experiment, to compare and contrast the Hanbury Brown–Twiss (HBT) effect for both 'bosonic' and 'fermionic' particles.

Although the original HBT effect can easily be understood within the framework of classical physics, explaining it in quantum-mechanical terms is more tricky. It requires acknowledging that photons are particles of integer spin, or bosons. These particles are far more gregarious than their fermion (half-integer-spin) cousins, and the bunching phenomenon can be described as the result of constructive interference of the quantum-mechanical probability amplitudes of two (bosonic) photons reaching the detectors. This explanation led Roy Glauber[3] to formulate modern photon-counting theory within the framework of quantum electrodynamics, the quantum field theory of the electromagnetic force. The result was the birth of modern quantum optics, an achievement crowned by a Nobel prize for Glauber in 2005.

Atoms of the helium isotope ^{4}He are also bosons, because they consist of a total of six half-integer-spin particles: four nucleons (two protons and two neutrons) and two orbiting electrons. Experiments with ultracold, metastable ^{4}He have not only famously allowed the observation of Bose–Einstein condensation (the phenomenon of many bosons all adopting the same quantum state), but have also opened the way to precise time-resolved and position-sensitive counting experiments in atomic systems. These helium atoms have a very long lifetime if unperturbed, but can be detected with almost perfect efficiency in microchannel plate and delay-line detectors.

The first direct observation of the bunching of ^{4}He atoms — the atomic HBT effect — came two years ago[4]. The same authors are part of the team that has now seen[2] the analogous effect in ^{3}He atoms, whose two electrons and three nucleons (only one neutron this time) make them fermions. Fermions obey the Pauli exclusion principle, so unlike bosons they do not like being in the same place at the same time. What Jeltes *et al.*[2] observe in the case of ^{3}He, therefore, is not the bunching characteristic of bosons, but 'antibunching' resulting from the destructive interference of the fermions' probability amplitudes.

The measurement of the HBT effect gives insight into the 'pair-correlation functions', a measure of the probability of finding two atoms at a certain distance apart, and therefore of how an atomic system is put together. Jeltes and colleagues' experiments were performed with dilute, practically non-interacting clouds of atoms in a state of thermal equilibrium. The

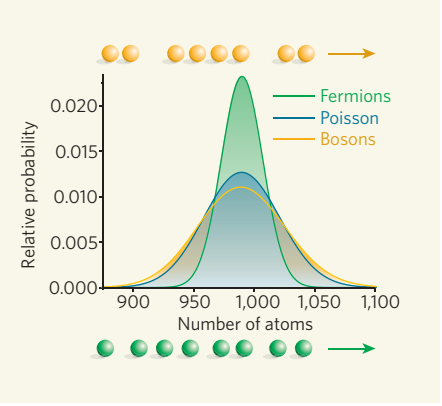
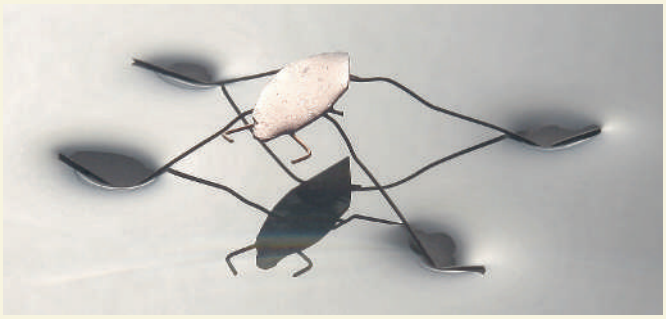


Figure 1 | Bunch, antibunch. Simulated distributions of the number of atoms detected in a certain time-window after 1,000 are released from an atomic trap at low temperature. If the arrival times of the atoms at the detectors were truly random, they would conform to a Poisson distribution (blue). In fact, bosons prefer to bunch, so in the time-window of any one detection event, significantly more or fewer than the mode number can arrive: the counting distribution (yellow) is broader. For fermions, the converse is true: they antibunch, arriving more regularly spaced than purely randomly, and so produce a taller, narrower counting distribution (green). (Figure courtesy S. Braungardt, U. Sen, A. Sen (De) & R. J. Glauber.)

SURFACE CHEMISTRY

Repellent legs



The glossy, water-repellent leaves of lotus plants (*Nelumbo nucifera* and *N. lutea*) have inspired many synthetic superhydrophobic surfaces. But the best of these are difficult to make, and can be very fragile. Steven Bell and colleagues now describe a simple method to prepare robust superhydrophobic surfaces of high quality (I. A. Larmour *et al. Angew. Chem. Int. Edn* doi:10.1002/anie.200604596; 2007).

The secret of lotus leaves' water-repellency is the double roughness of their surfaces — caused by the presence of nanohairs on microbumps — coupled with a waxy coating. The authors recreate this double roughness by coating metal substrates with a textured layer of another metal, simply by immersing them in a metal-salt solution. Scanning electron microscopy shows that the deposited metal forms flower-like structures (0.20 to 1 μm across) that are made up of smaller crystallites (about 60 to 200 nm in size), simulating the complexity of lotus leaf surfaces.

Dipping the substrates in a solution of a chemical surface-modifier, HDFT, supplies a monolayer of hydrophobic molecules. These molecules are highly fluorinated, just like the Teflon lining of non-stick frying-pans. The resulting surfaces show almost perfect super-hydrophobicity: a drop of water on a perfectly water-repellent surface forms a contact angle θ of 180°, and the surfaces produced by this method have θ values consistently greater than 170°.

The approach is so simple that it can be applied to metal objects of any reasonable size or shape. The authors again turned to nature for inspiration. Pond skaters (*Gerridae*) use superhydrophobic legs to walk on water. Bell and colleagues made a model pond skater from copper (pictured), with legs that had been treated with silver and HDFT. Despite having ten times the mass of a real pond skater, the metallic insect was able to rest comfortably on the surface of water.

The authors suggest that their method will aid research into superhydrophobic surfaces. This should hasten the arrival of practical applications, such as reducing turbulent flow in water-bearing pipes.

Andrew Mitchinson

ultimate goal is to measure the pair correlations, or perhaps higher-order correlations between more than two particles, for various strongly correlated, interacting quantum systems. Such measurements are technically demanding, but the authors show how a significant step can be made towards that goal by using an atomic lens. This is a laser that forces atoms away from its axis, 'defocusing' the atomic clouds, spreading them out in space and so significantly increasing the resolution for detecting the position of individual atoms.

The atomic-lensing method should allow the observation of, for instance, the antibunching that a one-dimensional gas of bosons undergoes because of 'fermionization' as a result of increased atom–atom repulsion in a confined space. Analysing the raw data obtained from atom detectors, one should also be able to extract the full atomic counting distribution. For non-interacting bosonic atoms in thermal equilibrium, this distribution should be broader than a Poisson distribution (Fig. 1); for non-interacting fermions, it should be narrower.

Direct counting of atoms at high resolution is so far possible only with metastable helium atoms, which limits the application of the method. Alternatives are in development. The pair correlations of a Bose–Einstein condensate that was split into two interfering parts was measured a few years ago[5]. The detection of single atoms passing through a high-quality optical cavity is possible, and was also used[6] to measure bosonic counting statistics and the bosonic HBT effect.

Another powerful method for measuring pair-correlation functions is noise interferometry. This is particularly useful for ultracold gases in an optical lattice[7], a perfect periodic potential made by interfering laser beams. When this optical potential is strong, bosonic atoms cannot tunnel from lattice site to lattice site. They form instead a 'Mott insulator' state with a fixed number of bosons per site. The images obtained when these bosons are released from the lattice are noisy and blurred, indicating a lack of phase correlation. Analysing the noise correlations in a sequence of such images has been used to assess, for example, the pair-correlation functions of the Mott-insulator state of bosonic rubidium-87 atoms[8]. A band insulator of polarized fermionic potassium-40 atoms in a lattice has also been constructed[9]. This is a state in which atoms completely fill the lowest energy band; as in the Mott insulator, there is no possibility of tunnelling, and no site-to-site phase coherence. Whereas noise interferometry in a bosonic Mott-insulator system produces a periodic sequence of peaks (Fig. 2) indicative of bunching, here it leads to a series of dips, equivalent to fermionic anti-bunching.

These noise interferometry investigations and atom-counting experiments such as those of Jeltes and colleagues[2] will continue to supply fascinating information on the physics of strongly correlated quantum many-body systems and their constituents. As our methods develop, so our prying into the private and social lives of particles will become ever more pervasive. ■

Maciej Lewenstein is at the Institució Catalana de Recerca i Estudis Avançats and the Institut

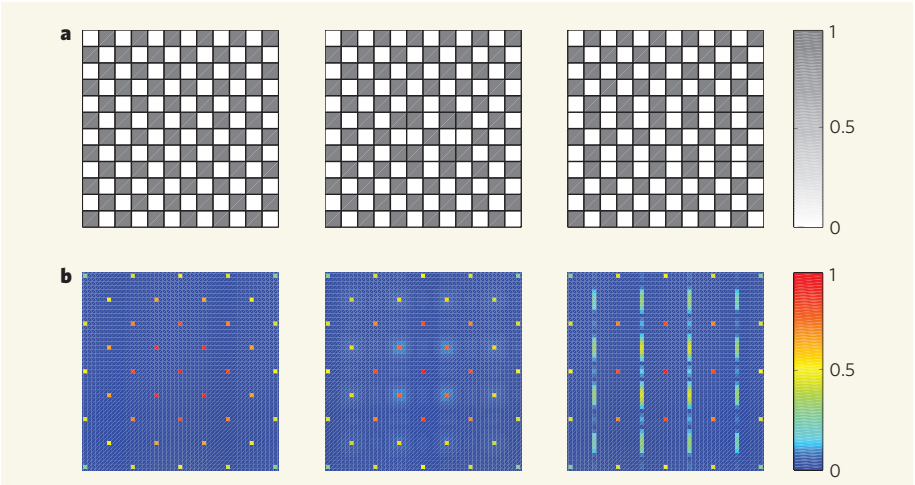


Figure 2 | Noise assessment. 'Noise interferometry' is a particularly efficient way of assessing spatial structures. **a**, A two-dimensional, dipolar Bose-gas Mott-insulator state held in an optical lattice, for example, ideally forms a chequer-board state of alternating filled and vacant sites at low temperatures and half filling[10] (left diagram; dark sites indicate presence of an atom). In practice, various kinds of defects occur (adjacent squares filled or unfilled; middle and right diagrams). **b**, Noise interferometry converts this spatial pattern into an easily identifiable interference signal, a characteristic series of peaks equivalent to a bunching behaviour.

de Ciències Fotòniques, 08869 Castelldefels (Barcelona), Spain.
e-mail: maciej.lewenstein@icfo.es

1. Hanbury Brown, R. & Twiss, R. Q. *Nature* **177,** 27–29 (1956).
2. Jeltes, T. *et al. Nature* **445,** 402–405 (2007).
3. Glauber, R. J. in *Quantum Optics and Electronics* (eds DeWitt, B., Blandin, C. & Cohen-Tannoudji, C.) 63–185 (Gordon & Breach, New York, 1965).
4. Schellekens, M. *et al. Science* **310,** 648–651 (2005).
5. Hellweg, D. *et al. Phys. Rev. Lett.* **91,** 010406 (2003).
6. Öttl, A., Ritter, S., Köhl, M. & Esslinger, T. *Phys. Rev. Lett.* **95,** 090404 (2005).
7. Altman, E., Demler, E. & Lukin, M. D. *Phys. Rev. A* **70,** 013603 (2004).
8. Fölling, S. *et al. Nature* **434,** 481–484 (2005).
9. Rom, T. *et al. Nature* **444,** 733–736 (2006).
10. Mennotti, C., Trefzger, C. & Lewenstein, M. preprint available at www.arxiv.org/cond-mat/0612498 (2006).

STRUCTURAL BIOLOGY

Pass the protein

Jean-François Trempe and Jane A. Endicott

Modifier proteins, such as ubiquitin, are passed sequentially between trios of enzymes, like batons in a relay race. Crystal structures suggest the mechanism of transfer between the first two enzymes.

Proteins are essential for all cellular processes, but their activities must be tightly controlled. One way of doing this is to covalently attach a small protein such as ubiquitin, or other ubiquitin-like proteins (UBLs)[1,2] such as NEDD8 and SUMO. For example, many proteins that control cell division must be degraded at precise points during the cell cycle. The initial step in this process is achieved by tagging the protein with ubiquitin.

Despite the variety of UBL functions, the enzymes that attach UBLs to proteins are remarkably similar, and share the same general mechanisms of action[3]. But how do these enzymes work? On page 394 of this issue, Huang *et al.*[4] describe the structure of an intermediate protein complex that forms during the process of attaching UBLs to proteins*. This structure gives valuable insights into the mechanism by which UBLs act in cell-signalling pathways.

UBLs become attached to proteins in a series of reactions catalysed by a trio of enzymes — described generically as E1, E2 and E3. The UBLs are first activated by E1, using energy derived from adenosine triphosphate (ATP) molecules. In this step, one end of the UBL becomes adenylated — that is, covalently attached to adenosine monophosphate — to generate a reactive intermediate that binds tightly to E1's adenylation active site (A-site). The UBL is then transferred to another active site within E1, the T-site, where it becomes attached to a cysteine amino acid. This relatively stable E1~UBL adduct transfers its UBL cargo to a cysteine in the next enzyme of the sequence, E2. The resulting E2~UBL adduct then interacts with the third component of the cascade, the E3 enzyme. This acts as a bridge between the E2~UBL adduct and the protein receiving the UBL, and thereby provides substrate specificity to the pathway.

What do we know about the molecular details of these reactions? The crystal structures of two E1–UBL complexes have been determined[5,6]. In each case, a single UBL molecule binds within a cleft from which the flexible UBL tail extends, so that the end of the tail is positioned close to the A-site of E1 (Fig. 1a). But although these structures suggest mechanisms by which an E1 enzyme can recognize and activate its particular UBL, they do not reveal how the activated UBL is transferred from the A-site to the T-site.

We also know something about how an E1 enzyme passes its UBL cargo to an E2. The E1 enzyme for NEDD8 consists of two proteins, called UBA3 and APPBP-1. The structure of a complex between the E2 enzyme of NEDD8 and a fragment of UBA3 has been determined[7]. Intriguingly, when this structure is superimposed on the structure of an entire E1–UBL complex, the E1 and E2 active-site cysteines are remote from each other — too far apart for a direct transfer of the UBL from one to the other. So how is this hand-over effected?

The answer lies in Huang and colleagues' crystal structure[4], which depicts a complex formed from two NEDD8 molecules, E1, E2 and an ATP molecule associated with a magnesium ion. In this structure, the E1 enzyme is folded into three domains that create a large central groove; the groove is divided into two clefts by a 'crossover' loop, which connects two of the domains. One NEDD8 molecule, here designated UBL(T), is covalently attached to the active cysteine of E1 at the T-site; the other NEDD8 molecule binds to the same part of the cleft that was occupied by the UBL in the singly loaded E1 structures described above[5,6]. On the basis of their structure, Huang *et al.* propose that a change in conformation of the protein complex allows the UBL to be passed from E1 to E2.

Central to this conformational switch is a region of UBA3 known as the ubiquitin-fold domain (UFD). In the previously determined structures that are loaded with just one UBL, this domain occupies the position of UBL(T), but in the doubly loaded structure[4] it is rotated 120° away from the T-site (Fig. 1b). The E2 enzyme binds to the UFD in this rotated position, so that the active cysteine of E2 lies enticingly close to the UBL-loaded cysteine of E1. Rather elegantly, this conformational change also alters the affinity of E1 for E2, creating an affinity switch: E1 doubly loaded with NEDD8 displays high affinity for E2, whereas singly loaded E1 does not.

But the story does not end there. As the E2 binding sites for E1 and E3 are mutually exclusive[8], the E2~UBL adduct must dissociate from E1 in order to associate with an E3. Does it jump, or is it pushed? The answer is possibly the latter. Following the discharge of UBL from the T-site, E1 reverts to its singly loaded

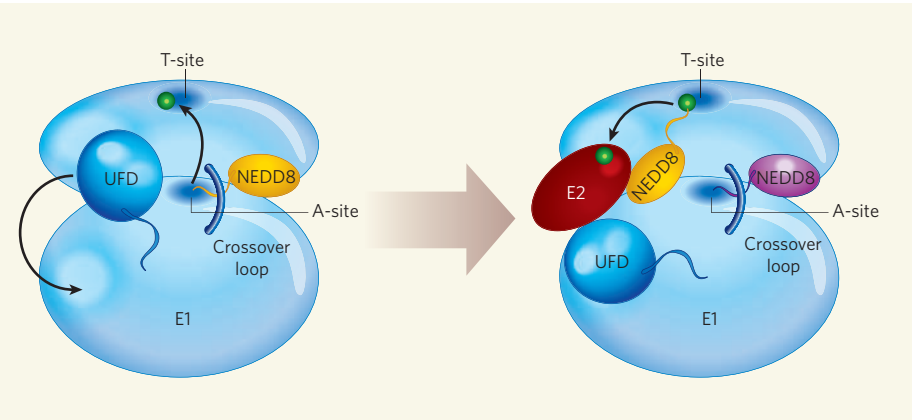


Figure 1 | A protein conformational switch. a, Before they become attached to a target protein, ubiquitin-like proteins such as NEDD8 are transferred from an E1 to an E2 enzyme. Crystal structures show that the tail of a NEDD8 protein (yellow) that is bound to E1 passes under a 'crossover' loop; the tail's end reacts with adenosine triphosphate (ATP) at the ATP-binding site of E1 (the A-site). The NEDD8 molecule is then transferred to a cysteine amino acid at the 'T-site', and a region of E1 known as the ubiquitin-fold domain (UFD) changes position. **b,** Huang *et al.*[4] propose that the combined conformational changes create a surface to which an E2 enzyme binds with high affinity. A second NEDD8 molecule (purple) now also binds to the A-site. The first NEDD8 molecule (yellow) is then transferred to a cysteine in the active site of the E2 enzyme. The positions of the active cysteines in E1 and E2 are marked by green circles.

*This article and the paper concerned[4] were published online on 14 January 2007.

NATURE|Vol 445|25 January 2007

conformation. This would cause the E2~UBL bound to E1 to clash with UBA3, promoting the departure of E2~UBL. This ingenious mechanism ensures an irreversible flow of UBL molecules from E1 via E2 to a substrate bound to E3.

Some mechanistic aspects of the E1–E2 transfer process remain to be addressed. The A-site and the T-site of E1 are separated by the prominent crossover loop of UBA3 (Fig. 1a). How, then, does the topologically challenged transfer of the UBL between these sites occur? Two mechanisms suggest themselves: either the adenylated tail of the UBL changes conformation and comes close to the active cysteine in the T-site, or the UBA3 catalytic domain undergoes a conformational change that allows this cysteine to reach the A-site. It might be possible to distinguish between these two mechanisms by determining the structure of an E1 in complex with an adenylated UBL.

Finally, an interesting parallel can be drawn between the mechanism proposed by Huang *et al.*[4] and the E2–E3-catalysed attachment of ubiquitin to its protein targets. A structural model of an E2–E3 complex has been proposed[9] in which there is a large gap between the active cysteine of E2 and the substrate that is bound to E3. Could the conformational change reported by Huang *et al.* be a model mechanism for other reactions in signalling pathways involving UBLs? ■

Jean-François Trempe and Jane A. Endicott are in the Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK.
e-mail: jane.endicott@biop.ox.ac.uk

1. Hershko, A. & Ciechanover, A. *Annu. Rev. Biochem.* **67,** 425–479 (1998).
2. Welchman, R. L., Gordon, C. & Mayer, R. J. *Nature Rev. Mol. Cell Biol.* **6,** 599–609 (2005).
3. Bohnsack, R. N. & Haas, A. L. *J. Biol. Chem.* **278,** 26823–26830 (2003).
4. Huang, D. T. *et al. Nature* **445,** 394–398 (2007).
5. Walden, H. *et al. Mol. Cell* **12,** 1427–1437 (2003).
6. Lois, L. M. & Lima, C. D. *EMBO J.* **24,** 439–451 (2005).
7. Huang, D. T. *et al. Mol. Cell* **17,** 341–350 (2005).
8. Eletr, Z. M., Huang, D. T., Duda, D. M., Schulman, B. A. & Kuhlman, B. *Nature Struct. Mol. Biol.* **12,** 933–934 (2005).
9. Zheng, N. *et al. Nature* **416,** 703–709 (2002).

PLANETARY SCIENCE

Inside Enceladus

John Spencer and David Grinspoon

Chemical analysis of a plume emanating from near the south pole of Enceladus indicates that the interior of this saturnian moon is hot. Could it have been hot enough for complex organic molecules to be made?

Tiny, icy Enceladus is, at a mere 500 kilometres in diameter, Saturn's sixth-largest moon. But thanks to the discovery in July 2005, by NASA's Cassini Saturn orbiter, of a remarkable water-rich plume jetting from warm fractures near Enceladus's south pole[1–3] (Fig. 1), the satellite has rivalled the giant moon Titan as a focus of attention in the Saturn system. In a paper to be published in *Icarus*, Matson *et al.*[4] suggest that the plume may be the external manifestation of a chemically rich, subsurface hydrothermal system that reaches temperatures of more than 200 °C.

This activity offers planetary scientists the first real possibility in the Solar System of studying cryovolcanism — volcano-like activity involving ices, rather than the molten silicates of earthly volcanoes — as it happens. The plume is also an unprecedented opportunity for direct analysis, with Cassini's instruments, of material that was in the interior of an icy satellite only minutes earlier. Of the issues to be resolved on Enceladus, one of the most interesting is that of the moon's internal temperature distribution. This controls its potential for sustaining liquid water, the chemistry that is possible, and even the potential suitability of the moon as a habitat for life.

The highest surface temperatures measured directly by Cassini's thermal infrared imaging instrument are about 145 K (–128 °C)[2], although a temperature of at least 180 K at the plume's source can be inferred indirectly. It has been suggested[3] that the comparable masses of gaseous and solid H_2O in the plume calculated from the Cassini data are produced most easily if the plume is generated by the boiling of liquid water into the vacuum of space, implying a temperature near the surface of at least 273 K (0 °C), the melting point of ice. But this startling conclusion is by no means certain. For one thing, the gaseous water in the plume seems to be moving much faster than most of the ice particles[1,3]. Thus, to sustain the ratio of ice particles to gas in the plume, which is measured to be near 1, much more gas must be produced than ice. Direct condensation from the vapour phase, without involvement of liquid water, may be sufficient to produce the ice particles at the observed rate. The plume might also be produced[5] by explosive release of gas from 'clathrate' water ice, which has other molecules trapped in its crystal lattice, at temperatures well below 0 °C.

Whatever Enceladus's near-surface temperatures, things are presumably warmer deeper down. A subsurface ocean of liquid water, as is now thought to exist under the icy crust of Jupiter's moon Europa, is also possible on Enceladus — although direct evidence for such a thing is so far lacking. Matson *et al.*[4],

50 & 100 YEARS AGO

50 YEARS AGO

Swifts in a Tower. By Dr. David Lack — The domestic life of the swift...was until lately almost unknown, because it nests in holes which are commonly inaccessible, on high buildings, and often too far into these to be reached in any event. This book, however, describes ten years of observation of nesting swifts 'from the inside'. The opportunity was offered by the colony in the ventilation holes in the tower of the University Museum at Oxford (associated in memory with the historic debate on evolution between Huxley and Wilberforce); it was ingeniously taken, by the insertion of nesting boxes with lids and glass panels to study the birds at closest quarters from within the tower. Many ornithologists and others have climbed the tall interior ladders to see the birds, which are unafraid of approach in these circumstances; but it is to many hours of patient observation by the author, and by his wife and other assistants, that we owe the wealth of information now presented in such attractive form. The results are of great interest, and are illustrated by remarkable electronic flash photographs by Mr H. N. Southern.

ALSO

Mr. Duncan Sandys, Minister of Housing and Local Government, has confirmed an order establishing the Gower Peninsula, Glamorgan, as an area of outstanding natural beauty... the first of its kind under the National Parks and Access to the Countryside Act, 1949.

From *Nature* 26 January 1957.

100 YEARS AGO

The pipe line conveying petroleum from Baku to the Black Sea has been completed. It is 550 miles long, and is capable of passing 400,000,000 gallons of oil yearly. Another important oil-pipe line has been built for transporting Texas and California petroleum across the Isthmus of Panama. It is 8 inches in diameter and fifty-one miles long.

From *Nature* 24 January 1907.

BIOGEOGRAPHY

Bounty beneath the Nullarbor

G. J. PRIDEAUX

Over a period of several hundred thousand years, many visitors dropped into Leaena's Breath cave beneath the Nullarbor plain in southern Australia but never left. The remains of these hapless animals, in this and two associated caves, constitute a palaeontological bounty for understanding past conditions in the region during the middle Pleistocene. The discoveries and their environmental context are described by Gavin Prideaux and colleagues elsewhere in this issue (G. J. Prideaux *et al. Nature* **445,** 422–425; 2007).

The small entrance to Leaena's Breath cave was the undoing of a large number of mammals and reptiles. They evidently fell through this hole, dropping some 20 metres to the cave floor. If they were not killed by their injuries, they later died of thirst. By far the commonest remains are fossils of various marsupials such as wombats, opossums and especially kangaroos; many species of these animals were previously not known, and many did not survive the Pleistocene. Prideaux *et al.* applied a battery of techniques to date the fossils and the layers in which they were buried. Their results produce ages ranging between 780,000 and 200,000 years ago.

The Nullarbor plain is vast and empty, and today appears as it is in this photograph: flat, dry, shrubby and almost treeless. From their analyses of isotope ratios in samples of herbivore tooth enamel, both ancient and modern, and the faunal composition, the authors conclude that in the past the Nullarbor had a more diverse flora, and a mixture of woods and shrubland that contained more plants with palatable leaves and fruits. But given that the species that did become extinct seem to have been as well adapted to dry conditions as those that did not, the authors also think the environment was as arid then as it is now.

Instead of invoking climate change, that common suspect, they argue that the best explanation for the different flora was an increased incidence of bushfires. The result is the impoverished, but more fire-resistant, vegetation of today.

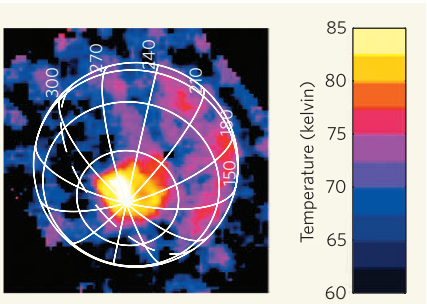
Tim Lincoln

however, find evidence in the composition of the gases ejected from Enceladus for interior temperatures even higher than those needed for liquid water. The gas composition, measured by Cassini's Ion and Neutral Mass Spectrometer (INMS)[6], is 91% water, 3.2% carbon dioxide, 4% nitrogen and 1.6% methane, with probable trace amounts of the organic gases acetylene and propane. Ammonia gas, NH_3, which has often been proposed as a significant component of icy satellite interiors[7] and an enabler of cryovolcanism (it is a potent antifreeze), is conspicuous by its absence.

Matson and colleagues address the origin of these non-water species by drawing analogies to the atmosphere of Titan, which is dominated by nitrogen gas, N_2. The European Space Agency's Huygens probe, carried by Cassini to Titan two years ago, measured a huge depletion of the argon isotope ^{36}Ar relative to N_2 in Titan's atmosphere[8]. The implication of this finding is that Titan is unlikely to have formed at temperatures low enough that N_2 could have been incorporated directly, as a pure ice or trapped in water-ice as a clathrate: ^{36}Ar, which has a similar volatility, would then have been incorporated into Titan too. It's more likely that Titan's nitrogen came in the form of NH_3, which can survive as a solid at higher temperatures that would drive off both N_2 and argon[9]. Ultraviolet photolysis in Titan's atmosphere is proposed to have dissociated the NH_3, later to yield the observed N_2.

By extension, Enceladus, which accreted from the same circum-saturnian nebula as Titan, is likely to have acquired NH_3, but not N_2. Thus the N_2 we see in the plume today is probably ultimately derived from NH_3. But how? In contrast to Titan, photolytic conversion is unlikely: Enceladus's feeble gravity could not have prevented the escape of any N_2 produced by photolysis of NH_3 on the surface, as could have happened on the larger Titan. Matson *et al.* instead suggest that the interior of Enceladus is warm enough for thermal decomposition of NH_3 to N_2 — a process that requires temperatures of at least 575 K, even in the presence of a catalyst.

At these temperatures, other interesting chemistry would be possible, if appropriate catalysts were available. Methane (CH_4) could be generated from carbon monoxide with the addition, say, of hydrogen from NH_3 decomposition — although it is also plausible that the methane seen on Enceladus is primordial. Higher-mass hydrocarbons such as the propane and acetylene tentatively detected by the INMS could also be produced, and there is the potential to generate many other, more complex organic molecules. Admittedly, we can't be sure when this high-temperature chemistry might have occurred: as Matson *et al.* point out, it is possible that most of the action was early in Enceladus's history, and that lower temperatures now prevail.



NASA/JPL/GSFC/SOUTHWEST RES. INST.

Figure 1 | Hot cracks. A map of the surface temperatures on Enceladus made by the Cassini Composite Infrared Spectrometer (CIRS) on 9 November 2006 shows the excess heat radiation from the fractures in the southern polar region. Although the average south polar temperature is only 85 K, the CIRS spectra show that small regions reach at least 145 K. Matson *et al.*[4] suggest much warmer temperatures at Enceladus's interior.

The inevitable question is whether life might have arisen in this warm, wet, chemically rich environment. Europa's ocean has long been a favoured potential oasis for life in the outer Solar System. But Europa's secrets are locked beneath kilometres of ice. Enceladus, by conveniently venting its guts into space where we can study them, gives us a far better opportunity to not just ask, but perhaps to answer, that enormous question.

Cassini is by no means finished with Enceladus. The fly-by of 2005 merely skirted the edge of the plume, and Cassini can analyse gas hundreds of times denser by flying closer to the plume source. That could yield much more precise constraints on the chemistry of its interior. The next close Enceladus fly-by will be in March 2008, and at least five more close fly-bys are likely in Cassini's extended mission, now being planned for the period from mid-2008 to mid-2010. If there is life there, or even complex prebiotic organic chemistry, these encounters will increase our chances of catching its chemical scent. Future missions to Enceladus, the possibility of which is now being studied, could provide more definitive answers. ■

John Spencer is at the Southwest Research Institute, 1050 Walnut Street, Boulder, Colorado 80302, USA. David Grinspoon is at the Denver Museum of Nature & Science, 2001 Colorado Boulevard, Denver, Colorado 80205, USA.
e-mails: spencer@boulder.swri.edu; david.grinspoon@dmns.org

1. Hansen, C. J. *et al. Science* **311,** 1422–1425 (2006).
2. Spencer, J. R. *et al. Science* **311,** 1401–1405 (2006).
3. Porco, C. C. *et al. Science* **311,** 1393–1401 (2006).
4. Matson, D. L., Castillo, J. C., Lunine, J. & Johnson, T. V. *Icarus* (in the press); doi:10.1016/j.icarus.2006.10.016 (2006).
5. Kieffer, S. W. *et al. Science* **314,** 1764–1766 (2006).
6. Waite, J. H. *et al. Science* **311,** 1419–1422 (2006).
7. Squyres, S. W., Reynolds, R. T. & Cassen, P. M. *Icarus* **53,** 319–331 (1983).
8. Niemann, H. B. *et al. Nature* **438,** 779–784 (2005).
9. Hersant, F. *et al. Icarus* (submitted).

Q&A

CELL BIOLOGY

Chromosome territories

Karen J. Meaburn and Tom Misteli

The natural habitat of eukaryotic genomes is the cell nucleus, where each chromosome is confined to a discrete region, referred to as a chromosome territory. This spatial organization is emerging as a crucial aspect of gene regulation and genome stability in health and disease.

What do chromosome territories look like?

The word 'chromosome' usually conjures up striking images of the dense, X-shaped entities seen during cell division. It is easy to forget that for most of the time chromosomes exist as unravelled structures and their arrangement is confined by the boundaries of the cell nucleus. We now know that each chromosome maintains its individuality during the cell cycle and occupies a spatially limited volume, known as a chromosome territory. Using fluorescent tags, these can be seen *in vivo* as roughly spherical domains of about 2 micrometres in diameter (Fig. 1a, b; Box 1, overleaf).

Do all cells have them?

Chromosome territories can only form in cells that have a nucleus (eukaryotic cells), and most higher eukaryotes are thought to have them. But some lower eukaryotes, such as the yeast *Saccharomyces cerevisiae*, lack chromosome territories and their chromosomes seem to be more loosely arranged.

Do the territories have an internal structure?

The interiors of chromosome territories are permeated by highly branched, interconnected networks of channels. These make the genome sequences deep inside accessible to regulatory factors such as gene activators and inhibitors (Fig. 1c). In addition, the structure of the DNA within chromosome territories is nonrandom, as the chromosome arms are mostly kept apart from each other and gene-rich chromosome regions are separated from gene-poor regions. This arrangement probably contributes to the structural organization of the chromosome, and might also help in regulating particular sets of genes in a coordinated manner.

So how does a chromosome fold up into this form?

This is not yet clear. Observations in plants suggest that the chromatin fibre — comprising the DNA and its associated proteins — forms large loops that are anchored to each other at their base (Fig. 1d). However, in higher eukaryotes such as mammals the fibre seems rather to be folded into distinct megabase-sized domains that are linked to one another (Fig. 1e). Each of these domains might represent a functional unit, because their replication is coordinated and maintained in consecutive cell cycles. A hybrid model in which smaller loops emanate from a central chromosome core has also been suggested (Fig. 1f). In addition, there is debate about how chromatin fibres are organized at the surface of the chromosome territory. Some observations suggest that giant chromatin loops protrude from the chromosome territory and intermingle extensively with fibres from neighbouring chromosomes, whereas other studies seem to show that abutting chromosomes don't mix much. Emerging high-resolution light-microscopy methods should settle this issue soon.

Are the territories arranged in particular patterns within the nucleus?

Remarkably, yes. In lower eukaryotes such as plants and flies chromosomes tend to be polarized, with the ends of the arms (telomeres) on one side of the cell nucleus and the point at which the two arms meet (the centromere) on the opposite side. In mammalian cells, however, chromosome arrangement is more complex. Even so, each chromosome can be assigned a preferential position relative to the nuclear centre, with particular chromosomes tending to be at the nuclear interior and others at the edge (Fig. 2a, overleaf). This preferential radial arrangement also, of course, gives rise to preferred clusters of neighbouring chromosomes.

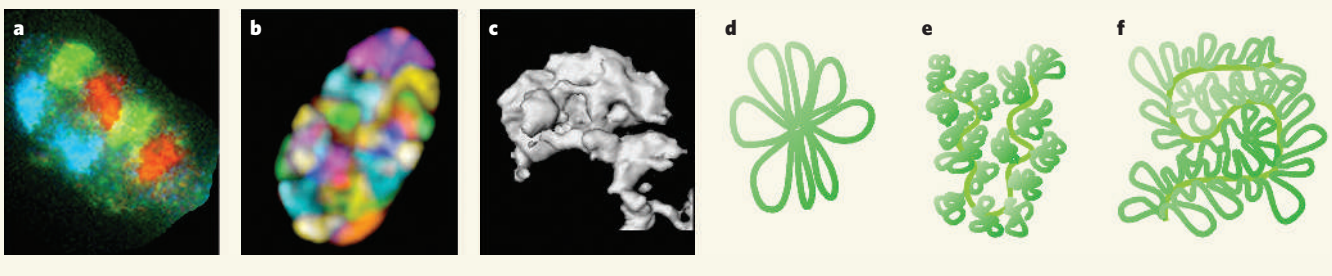
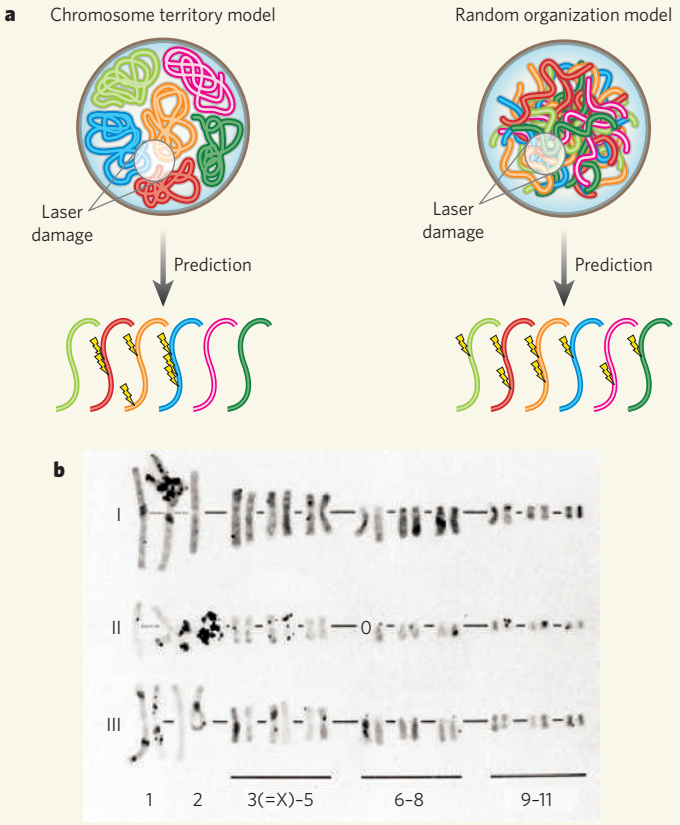


Figure 1 | Spatial organization in the nucleus. **a**, Territories can be visualized using chromosome-specific fluorescent probes. In this mouse liver nucleus, chromosome 12 (red), chromosome 14 (green) and chromosome 15 (blue) were painted. **b**, Modern imaging techniques allow all chromosomes in a cell to be visualized simultaneously, as seen here in a human fibroblast. **c**, Chromosome territories are not solid entities, and their interior is permeated by a network of nucleoplasmic channels. **d**–**f**, Higher-order organization of chromosome territories. In plants, the chromatin fibre forms a rosette-like structure (**d**), whereas in higher eukaryotes chromatin forms interconnected megabase-sized domains (**e**). Other models suggest looping of the fibre from a central backbone (**f**). (Panel **a** courtesy of L. Parada; panels **b** and **c** courtesy of T. Cremer.)

Box 1 | The discovery of chromosome territories



At the turn of the twentieth century, Carl Rabl and Theodor Boveri proposed that each chromosome maintains its individuality during the cell cycle, and Boveri explained this behaviour in terms of 'chromosome territories'.

The existence of chromosome territories was demonstrated experimentally during the early 1980s in pioneering microlaser experiments by the brothers Thomas and Christoph Cremer. They used a microlaser to induce local genome damage, and predicted that inflicting DNA damage within a small volume of the nucleus would yield different results depending on how chromosomes were arranged. If chromosomes occupied distinct territories (**a**, left panel), localized damage would affect only a small subset of chromosomes, whereas if the chromatin fibres of each chromosome were randomly distributed throughout the nucleus (**a**, right panel), many chromosomes would be damaged. **b**, Three sets (I–III) of hamster chromosomes after laser damage. Only a subset of the chromosomes was damaged, as indicated by the black grains of radioactivity most prominently seen on chromosomes 1 and 2. This demonstrates the existence of chromosome territories. (Panel **b** was reprinted with permission from C. Zorn *et al. Exp. Cell Res.* **124,** 111–119; 1979.) **K.J.M. & T.M.**

Is the arrangement always the same, then?

First of all, the patterns are probabilistic, rather than absolute; so although a chromosome may have a preferred average position in a cell population, the location of the chromosome in individual cells within that population can vary greatly. Even the two copies of the same chromosome within the same nucleus often occupy distinct positions and have different immediate neighbours.

Chromosome arrangements are also specific to the cell and tissue type, and can change during processes such as differentiation and development. For example, during differentiation of immune T cells, mouse chromosome 6 moves from an internal position to the nuclear periphery.

The precise physiological relevance of chromosome positioning is currently unclear. However, its significance is hinted at by the fact that there is similarity in chromosome-position patterns among cell types that share common developmental pathways and by the observation that chromosome positions in a given cell type are evolutionarily conserved. For example, in human lymphocyte cells, chromosomes 18 and 19 tend to occupy a peripheral and an internal position, respectively — as does the corresponding genetic material in Old World monkeys.

Why have all this organization?

The nonrandom organization of the genome allows functional compartmentalization of the nuclear space. At the simplest level, active and inactive genome regions can be separated from each other, possibly to enhance the efficiency of gene expression or repression. Such compartmentalization might also act in more subtle ways to bring co-regulated genes into physical proximity to coordinate their activities. For instance, in eukaryotes, the genes encoding ribosomal RNAs tend to cluster together in an organelle inside the nucleus known as the nucleolus. In addition, observations made in blood cells suggest that during differentiation co-regulated genes are recruited to shared regions of gene expression upon activation.

So, how do chromosomes find their place in the nucleus?

We don't know. Chromosomes are physically separated during cell division, but they tend to settle back into similar relative positions in the daughter cells, and then they remain stable throughout most of the cell cycle. So there must be some molecular mechanism that establishes and maintains the chromosomes' positions. The radial positioning of chromosomes has been related to either the chromosome gene density or the amount of DNA they contain, depending on cell type and proliferation status. But these cannot be the only factors involved, because the arrangement changes during differentiation and proliferation, when gene density and chromosome size remain constant.

What are the mechanisms of chromosome positioning?

There are two fundamentally different possibilities. It may be that chromosome positions are determined through their association with immobile nuclear elements — possibly a nuclear scaffold similar to the molecular structures that support and organize the cell's cytoplasm. Although such anchoring may explain chromosome immobility and stability during the cell cycle, it cannot account for nonrandom positioning unless there is some sort of tethering mechanism that is specific to each chromosome and also encodes positioning information.

An attractive alternative is a self-organization model in which the position of each chromosome is largely determined by the overall activity of all of its genes; that is, the number and pattern of active and silent genes on a given chromosome. The idea here is that the expression status of a genome region affects local chromatin structure, with inactive regions being more condensed (heterochromatin) and highly active ones decondensed (euchromatin). Depending on the degree of genome activity and the linear distribution of active and

inactive regions on a chromosome, different chromosomes are expected to have distinct overall physical properties. These might, in turn, determine their likelihood of interacting with each other and so affect their relative arrangement. This model explains the observed clustering of functionally equivalent chromosomal regions, such as the heterochromatic centromeres. It also explains the documented tissue-specificity of chromosome patterns, because chromosomes in different tissues express distinct subsets of genes.

Does nuclear position matter as far as individual genes are concerned?

The position within the nucleus of several genes has been linked to their activity. In yeast, certain genes relocate to the nuclear periphery after activation. In mammalian cells, some genes relocate from the periphery towards the interior once they have been switched on. Similarly, a number of highly transcribed gene-dense regions — for example the MHC class II locus — are expelled from the chromosome territory once they are activated by formation of a large chromatin loop. On the other hand, many other genes either do not change position upon activation or move in the opposite direction, and active genes are found both at the surface of chromosome territories and deeply buried within the territory.

So there isn't a simple rule that applies to all genes?

No. It seems that the actual position of a gene in the cell nucleus is not essential to its function. It is more likely that positioning contributes to optimizing gene activity. Indeed, in yeast the association between a gene and the nuclear periphery does not determine whether or not the gene is active *per se*, but seems instead merely to modulate a gene's expression to optimize it.

What about the position of genes relative to each other — does that affect how they function?

The relative positioning of DNA sequence elements to one another and their physical interaction is emerging as highly significant for several cellular functions (Fig. 2b). In particular, changing the relative arrangement of genome regions can bring regulatory regions into proximity with otherwise distant genes to control their function. It has long been known from observations in the fruitfly that placing a gene near a block of heterochromatin represses expression of that gene. Recent observations suggest that gene loci may also be controlled in a similar manner by interactions with particular regulatory regions on other chromosomes. The most striking example of such regulation comes from odorant receptor genes. Each mouse olfactory neuron expresses only one of more than 1,000 odorant receptors. Which gene is expressed is determined when

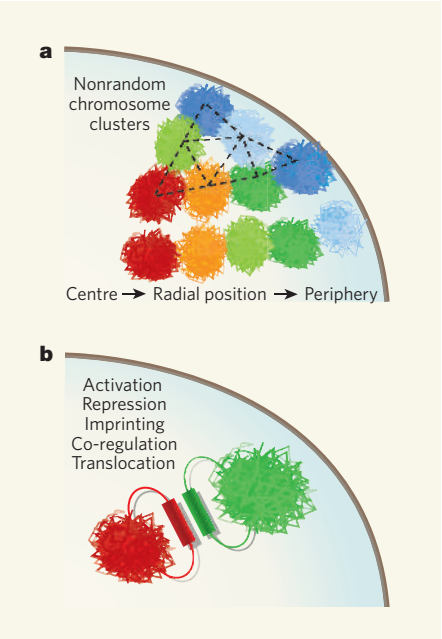


Figure 2 | Arrangement of chromosome territories. a, Chromosomes occupy nonrandom radial positions relative to the centre of the nucleus. Red and orange chromosomes are preferentially internal, green chromosomes intermediate and blue chromosomes peripheral. Positioning patterns are probabilistic, and preferential radial positions lead to nonrandom clusters of chromosomes. **b**, Changing the relative arrangement of genome regions (coloured rectangles) to bring them into close proximity is functionally relevant for gene activity, and for the formation of chromosomal translocations.

a regulatory element on one chromosome associates with an odorant receptor gene on a different chromosome to selectively activate it. Similar interchromosomal interactions may be involved in differentiation-specific gene activation in immune T cells and in gene imprinting — the process whereby the maternal or paternal copy of certain genes is permanently inactivated in a cell.

Are there any other consequences of the relative arrangement of genome regions?

Arrangement can affect genome stability, particularly the formation of cancer-promoting chromosome translocations. Such translocations occur when double-strand breaks in neighbouring chromosomes are not rapidly and correctly repaired, and the broken ends from different chromosomes are mistakenly joined. It turns out that chromosomes that are preferentially close to one another seem to undergo translocations more readily than those that are farther apart.

Is any of this of practical interest?

Apart from providing insight into genome function, it might be possible to exploit the positioning patterns of genes and chromosomes in the near future for diagnostic purposes. Any genome region that changes its nuclear position as a result of aberrant gene expression may be indicative of a disease process. For example, diseased cells could be identified by detecting changes in the positions of marker genes brought about by their misregulation. The advantage of such positional analysis would be the detection of aberrant cells in the context of intact tissue and in primary biopsy samples, without the need for cell culturing. In addition, because changes in positioning are an initial step in gene activation, relocalization of certain regions may permit early diagnosis of a disease. Such methods might be particularly useful for the analysis of solid tumours, which are mostly refractory to routine diagnostic chromosome analysis.

And what of the future?

This is a truly exciting field for many reasons. Identifying the rules and mechanisms that determine how genomes and chromosomes are spatially organized, and how their organization changes during physiological processes, is a logical continuation of our ongoing exploration of genome sequences. Without elucidating the cell biology of genomes, we will not understand how genomes function in intact living cells or how these functions go awry in disease.

So far, much of the work in this field has been descriptive and correlative. But these early mapping studies have provided the necessary framework to begin to design experiments that test the role of genome positioning in gene expression. The great aspiration for the field now is to identify the molecular mechanisms responsible for the repositioning of single genes, genome regions and whole chromosomes within the nuclear space, and to determine how these mechanisms respond and contribute to physiological cues, such as stimulation through signalling pathways. Such insight will contribute greatly to a full appreciation of how the one-dimensional DNA sequence gives rise to the multi-dimensional complexity of the gene-transcriptional networks that determine all aspects of life. ■

Karen J. Meaburn and Tom Misteli are at the National Cancer Institute, NIH, Bethesda, Maryland 20892, USA.
e-mail: mistelit@mail.nih.gov

FURTHER READING
Cremer, T. & Cremer, C. *Nature Rev. Genet.* **2,** 292–301 (2001).
Foster, H. A. & Bridger, J. M. *Chromosoma* **114,** 212–229 (2005).
Kosak, S. T. & Groudine, M. *Science* **306,** 644–647 (2004).
Meaburn, K. J., Misteli, T. & Soutoglou, E. *Semin. Cancer Biol.* **17,** 80–90 (2006).
Misteli, T. *Cell* **119,** 153–156 (2004).
Parada, L. & Misteli, T. *Trends Cell Biol.* **12,** 425–432 (2002).
Taddei, A., Hediger, F., Neumann, F. R. & Gasser, S. M. *Annu. Rev. Genet.* **38,** 305–345 (2004).

PROGRESS

Empirical fitness landscapes reveal accessible evolutionary paths

Frank J. Poelwijk[1]*, Daniel J. Kiviet[1]*, Daniel M. Weinreich[2]† & Sander J. Tans[1]

When attempting to understand evolution, we traditionally rely on analysing evolutionary outcomes, despite the fact that unseen intermediates determine its course. A handful of recent studies has begun to explore these intermediate evolutionary forms, which can be reconstructed in the laboratory. With this first view on empirical evolutionary landscapes, we can now finally start asking why particular evolutionary paths are taken.

Evolutionary intermediates represented a central preoccupation for Darwin in his case for the theory of evolution. He remarked, for example: '…why, if species have descended from other species by insensibly fine gradations, do we not everywhere see innumerable transitional forms?'[1]. Although Darwin developed a convincing rationale for their absence, he did realize that the lack of intermediates as proof leaves room for criticism. He noted, for instance: 'If it could be demonstrated that any complex organ existed which could not possibly have been formed by numerous, successive, slight modifications, my theory would absolutely break down.'[1]. Indeed, in their opposition to evolution, the proponents of 'intelligent design' have seized on our current ignorance of intermediates.

Building on earlier ideas[2–4], an approach has recently been developed to explore the step-by-step evolution of molecular functions. The central innovation is that all molecular intermediates along multiple putative pathways are explicitly reconstructed. Together with a phenotypic characterization of each intermediate, one can determine whether paths towards a certain novel function are accessible by natural selection. Although others have reconstructed and characterized phylogenetically ancestral forms of proteins[4–7], here the focus is on fitness landscapes[8] in which multiple mutational trajectories can be compared. Fitness landscapes have been widely studied on a theoretical level (see refs 9–13 for example), but one can now obtain a glimpse of actual biological landscapes. This view finally allows us to ask which particular evolutionary paths are taken and why. In particular, to what extent do biomolecular properties constrain evolution? Does it matter in which order mutations occur? Are fitness landscapes rugged, with many local optima acting as evolutionary dead-ends, or are they smooth? Is neutral genetic drift essential for a new trait to emerge?

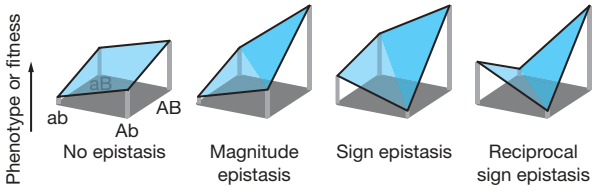
When examining the molecular underpinnings of the evolution of new traits, we distinguish two elementary cases. First, we discuss a single mutable component such as an enzyme. Second, we look at molecular interactions involving two or more mutable components, which is typical for regulatory evolution. The specific features of this broad range of molecular systems will be discussed using the notions of epistasis and fitness landscapes, which we will explain and relate to each other (Box 1 and Fig. 1).

The tentative picture emerging from the new results is one that emphasizes the possibilities of continuous optimization by positive selection. Although evolution was clearly constrained, as illustrated by many inaccessible evolutionary paths, the studies also revealed alternative accessible routes: a succession of viable intermediates exhibiting incremental performance increases. Although these findings do not address whether natural evolution proceeds in the presence or absence of selection, they do show that neutral genetic drift is not essential in the cases studied. We note that the presented approach starts with naturally occurring sequences, which are

Box 1 | Epistasis and the accessibility of mutational paths

Epistasis means that the phenotypic consequences of a mutation depend on the genetic background (genetic sequence) in which it occurs. In the Box figure we distinguish four cases that illustrate paths composed of two mutations, from the initial sequence 'ab' towards the optimum at 'AB'. When there is no epistasis, mutation 'a' to 'A' yields the same fitness effect for different genetic backgrounds ('b' or 'B'), while for magnitude epistasis the fitness effect differs in magnitude, but not in sign. For sign epistasis, the sign of the fitness effect changes. Finally, such a change in sign of the fitness effect can occur for both mutations, which we here term reciprocal sign epistasis.

These distinctions are crucial in the context of selection. Mutations exhibiting magnitude epistasis or no epistasis are always favoured (or disfavoured), regardless of the genetic background in which they appear. In contrast, mutations exhibiting sign epistasis may be rejected by natural selection, even if they are eventually required to increase fitness. In other words, some paths to the optimum contain fitness decreases, while other paths are monotonically increasing. When all paths between two sequences contain fitness decreases, there are two or more distinct peaks. The presence of multiple peaks indicates reciprocal sign epistasis, and may cause severe frustration of evolution (Fig. 1b). Indeed, reciprocal sign epistasis is a necessary condition for multiple peaks, although it does not guarantee it: the two optima in the diagram may be connected by a fitness-increasing path involving mutations in a third site.



[1]FOM Institute AMOLF, Kruislaan 407, 1098 SJ, Amsterdam, The Netherlands. [2]Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Avenue, Cambridge, Massachusetts 02138, USA. †Present address: Department of Ecology and Evolutionary Biology, and Center for Computational Molecular Biology, Brown University, Providence, Rhode Island 02192, USA.
*These authors contributed equally to this work.

themselves the product of evolution, and may therefore yield a biased sample of trajectories. Whether the conclusions are general or not, and whether they break down when the evolved feature becomes more complex, can only be determined through future studies.

Enzyme evolution

When a well-adapted organism is challenged by a new environment, an existing gene may perform suboptimally. One of the most basic questions one may then ask is: when mutating step-by-step from the suboptimal to an optimal allele, are all possible trajectories selectively accessible? This question depends critically on the stepwise changes in performance, or in fitness, which are governed by unknown physical and chemical properties at the molecular level. When all mutations along all paths yield a fitness improvement, evolution can rapidly proceed in a straightforward incremental darwinian fashion. In this case, the fitness landscape can be portrayed by a single smooth peak (Fig. 1a).

Whether this picture is realistic was investigated for the adaptation of bacterial β-lactamase to the novel antibiotic cefotaxime[14]. The central step was to reconstruct and measure all likely intermediates, allowing a systematic study of all possible trajectories. The intermediate sequences can be easily identified, because the (five) mutations that control the cefotaxime resistance phenotype are known, resulting in $2^5 = 32$ possible mutants. The order in which the mutations are fixed can of course be different, giving rise to $5! = 120$ possible direct trajectories between the start and end sequences.

The trajectory analysis showed that the fitness landscape is not as simple as depicted in Fig. 1a. A majority of the pathways towards maximum cefotaxime resistance actually shows a dip in fitness (see yellow path in Fig. 1b), or contain selectively neutral steps (as in Fig. 1c), resulting in much smaller chances of being followed by natural selection[12,15]. For 18 paths however, each step appeared to confer a resistance increase, making these trajectories accessible to darwinian selection. The part of the fitness landscape mapped out in this manner therefore does appear to have a single peak, but one that contains depressions and plateaus on its slopes. We stress that such three-dimensional analogies, while useful for conveying basic characteristics, do not rigorously represent the many direct trajectories existing between two alleles. Also note that there may be additional paths that contain detours, involving other mutations that are eventually reverted[16] (Fig. 1d).

Interestingly, some mutations yielded either a resistance increase or decrease, depending on the preceding mutations. This phenomenon, called sign epistasis[13] (see Box), is both a necessary and sufficient condition for the fitness landscape to contain inaccessible paths to an optimum[13]. Some cases of sign epistasis could be understood in terms of competing molecular mechanisms. For instance, a first mutation in the wild-type enzyme increased the resistance by enhancing the catalytic rate, even though it also lowered the thermodynamic stability. This loss of stability was repaired by a second mutation, thereby further increasing the resistance. In contrast, when this 'stabilizing' mutation occurred first in the wild-type enzyme, the resistance was reduced. Such back and forth balancing between structural and functional benefits might well be a more general evolutionary mechanism[17,18].

In a second study[19], the connection between fitness landscape and underlying molecular properties has been explored for the evolution of isopropylmalate dehydrogenase (IMDH, Fig. 2a), an enzyme that is involved in the biosynthesis of leucine. As in the previous study, a set of mutational intermediates between different functions were characterized. Here the mutations changed the cofactor binding affinity of IMDH. *In vitro* measurements of enzyme activity did not show epistasis: each mutation gave a fixed catalytic improvement, which was independent of the order in which they occurred. Thus, the 'enzyme activity' landscape is single-peaked.

The story becomes more complete with the following elements. First, the study also considered evolutionary paths from the suboptimal cofactor NADP to the normal cofactor NAD[20]. Second, selection does not act directly on enzyme activity, but rather on the fitness of an organism. As fitness is typically nonlinear in enzyme activity, epistasis is introduced. Therefore, the IMDH mutants were also evaluated *in vivo*, providing a direct measurement of the fitness effect of a mutation. The resulting fitness landscape was shown to contain a depression or valley, rendering the trajectories that pass through it

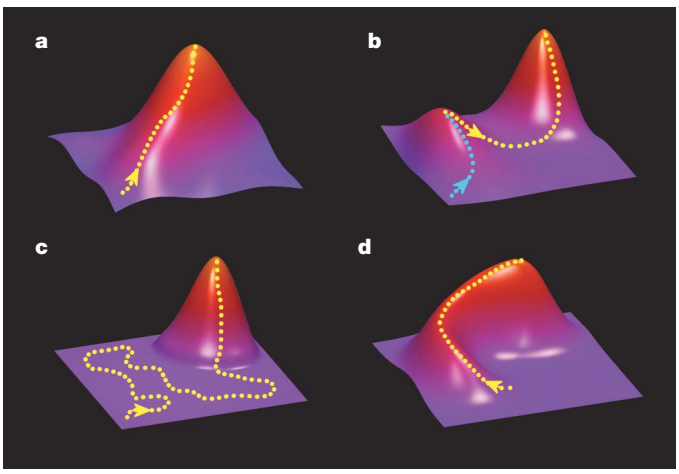


Figure 1 | Schematic representations of fitness landscape features. Fitness is shown as a function of sequence: the dotted lines are mutational paths to higher fitness. **a**, Single smooth peak. All direct paths to the top are increasing in fitness. **b**, Rugged landscape with multiple peaks. The yellow path has a fitness decrease that drastically lowers its evolutionary probability. Along the blue path selection leads in the wrong direction to an evolutionary trap[16]. **c**, Neutral landscape. When neutral mutations are essential, evolutionary probabilities are low[12,15]. **d**, Detour landscape. The occurrence of paths where mutations are reverted[16] shows that sequence analysis may fail to show mutations that are essential to the evolutionary history.

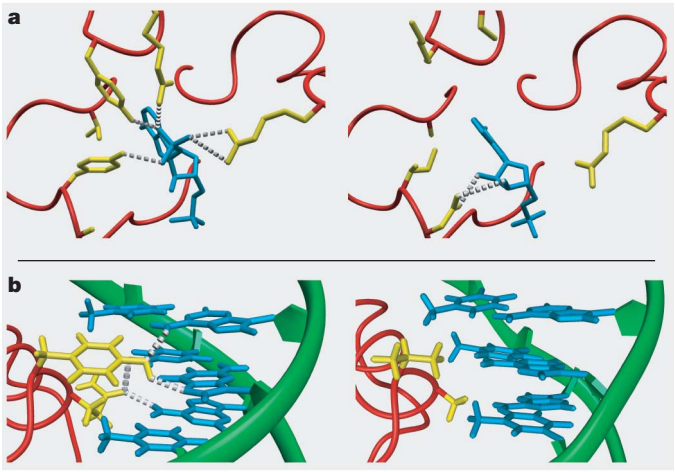


Figure 2 | Molecular structures in different evolutionary forms. Main chains are shown in red, key residues in yellow, the DNA backbone in green, key DNA bases or cofactor in blue, and hydrogen bonds as dashed lines. **a**, The left panel shows wild-type *E. coli* isocitrate dehydrogenase[34] (IDH), which is structurally similar to IMDH, with NADP as cofactor. The right panel shows an engineered IDH form with NAD as cofactor[35]. **b**, The left panel shows a wild-type *E. coli lac* repressor and operator[36]. The right panel shows a *lac* repressor and operator variant, with mutations mimicking the *gal* system[37]. Binding is tight and specific (despite the absence of hydrogen bonds): these variants bind wild-type partners poorly. Figures prepared with MOLMOL[38].

selectively inaccessible. There is an intuitive rationale for a valley here: when the recognition of NADP is reduced, the fitness first decreases, before it rises again when NAD recognition is built up. Interestingly however, some trajectories also exist that avoid the valley by simultaneously increasing NAD, and decreasing NADP recognition. In the end, the genotype–fitness landscape has a single peak, but one that includes a depression on its slope.

Evolution of molecular interactions

The evolutionary puzzle becomes more complex at a higher level of cellular organization. In the web of regulatory interactions between ligands, proteins and DNA, the components are strongly interdependent, which might suggest that their evolution is severely constrained. The evolution of molecular recognition has recently been explored by two studies, which also used experimentally reconstructed intermediates. The first examined hormone detection by steroid receptors in the basal vertebrates (Fig. 3a)[21]. The second[16] looked at the adaptation of repressor–operator binding, in a large evolutionary landscape based on published mutation data for the *Escherichia coli lac* system[22] (Figs 2b and 3b). For both studies, the molecular interactions may be thought of as a key fitting a lock. The unifying question is: can a new lock and matching key be formed taking just one mutational step at a time? The adaptation of these components presents a dilemma: if the lock is modified first, the intermediate is not viable because the old key does not fit, and vice versa.

From the evolution of the interactions in the two systems (Fig. 3), some interesting parallels are apparent. Both studies start with a duplication event yielding two locks and keys, and then ask how specific interactions can be obtained during mutational divergence. Specificity is clearly vital: two partners must recognize each other, but not recognizing other components is just as important. A major evolutionary challenge is therefore to decrease unwanted interactions, while maintaining desired interactions. Without specific hormone recognition, cortisol regulation of vertebrate metabolism, inflammation and immunity would be perturbed by varying levels of aldosterone, which controls electrolyte homeostasis. Similarly, specific recognition in the *lac* family of repressors allows *E. coli* to consume a wide array of sugars, without the burden of producing many unused metabolic enzymes.

Surprisingly, these studies again show that new interactions can evolve in a step-by-step darwinian fashion, despite the mismatching intermediates problem sketched above. In the hormone receptor case, this predicament is overcome by a molecular version of a master key: a putative ancestral ligand, 11-deoxycorticosterone, was found to activate all receptors (ancestral and present-day), allowing the mutational intermediates to remain functional even while the receptors diverged (Fig. 3a). The capability to synthesize aldosterone evolved later, finally providing a specific hormone that is recognized by just one of the two receptors. An existing receptor was thus recruited into a new role, as a binding partner to aldosterone, in a process that was termed 'molecular exploitation'. Sign epistasis was again observed: an initial mutation drastically lowered the response to all substrates, but after another mutation, the same mutation improved cortisol response while decreasing the aldosterone response. Thus, just as in the β-lactamase and IMDH cases, at least one selectively accessible evolutionary pathway existed.

In the evolution of the *lac* system, a similar mechanism using a 'master' repressor or operator was not observed. This is illustrated by the transient loss in affinity during the adaptation from one tight repressor–operator pair (IM–TG) to another (IK–AC); see Fig. 3b. Between some alleles, all connecting paths transiently reduced the affinity, indicating the presence of multiple peaks in the affinity landscape, which contrasts with the single-peaked landscapes of β-lactamase and IMDH. Multiple peaks indicate a severe kind of sign epistasis, which we here term reciprocal sign epistasis (see Box 1). Reciprocal sign epistasis can be intuitively understood for molecular interactions: mutating one binding partner will probably only benefit a new interaction if the other binding partner is mutated first, and vice versa. Interestingly, this means that although sign epistasis does introduce landscape ruggedness and thus perturbs the adaptive search, it can also be valuable because it enables multiple independent lock–key combinations.

If the *lac* repressor–operator affinity landscape is rugged and multi-peaked, how can new recognition evolve in a step-by-step manner? The answer lies in the fact that selection does not act on a single interaction. Instead, multiple interactions in a network determine the regulation, and ultimately organismal fitness. In the *lac* case, deteriorations in one interaction were offset by improvements in another. For example, initial mutations in one repressor duplicate were bad for binding to its designated operator, but good for relieving an undesired cross-interaction (Fig. 3b). These results substantiate the suggestion that network robustness[23] may promote evolvability[24,25]. The observed compensations yielded a smoothened fitness landscape, making the new interactions selectively accessible. In fact,

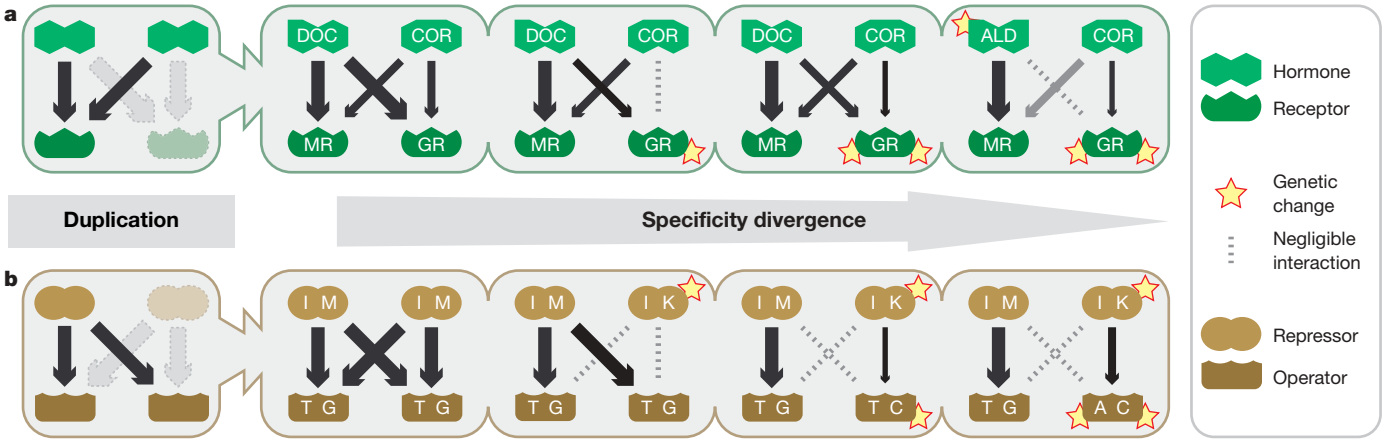


Figure 3 | Evolution of molecular interactions based on reconstructed intermediates. Arrow thickness denotes measured interaction strengths. DOC, 11-deoxycorticosterone; COR, cortisol; MR, mineralocorticoid receptor; GR, glucocorticoid receptor; ALD, aldosterone. **a**, Pathway towards independent steroid receptors after duplication, via intermediate receptors that remained sensitive to their ligands[21]. A changed mutation order produced a non-sensitive intermediate, making that path inaccessible. The grey arrow indicates that cortisol is absent in MR-expressing tissues. **b**, Pathway towards independent repressor–operator pairs following duplication, taking single-mutation steps without decreases in network performance. Many paths were compared in a landscape based on over 1,000 *lac* mutants[22], covering all substitutions on all key base pairs. For simplicity, the repressor dimer and two operator half-sites are not drawn.

because compensation within biochemical networks is ubiquitously observed[26], we expect that evolution by network compensation constitutes a general mode of regulatory adaptation, molecular interdependence notwithstanding.

Outlook

The experimental reconstruction of evolutionary intermediates and putative pathways has provided an exciting first look at molecular adaptive landscapes. Although numerous paths appear to be selectively inaccessible, accessible pathways are generally also available. Importantly, various alternative types of fitness landscapes were not observed. The landscapes could have been so rugged and multi-peaked, that accessible paths to optima would not exist, thus requiring, for instance, two or more simultaneous mutations, larger genetic modifications through recombination, or periods of relaxed selection. We have put forward various mechanisms that can reduce landscape ruggedness and improve evolvability. These include the interplay between protein function and stability[14,19], the exploitation of existing molecules into new roles[21], and compensation within biochemical networks[16].

That only a few paths are favoured also implies that evolution might be more reproducible than is commonly perceived, or even be predictable. It is important to note that evolutionary speed and predictability are not determined only by molecular constraints, but also by population dynamics. Population dynamics still presents many open questions, in particular in the context of regulatory evolution and varying environments. The situation in which environmental fluctuations are fast relative to selection timescales has been explored in the repressor divergence study[16]. Recent theoretical considerations[27,28] may provide promising approaches to address these questions more generally.

The molecular systems interrogated so far represent only a start, but one with great potential to spark further exploration. The analysis of intermediates is generally applicable, which makes finding new candidate systems not difficult. Mutational paths could also be revealed using the directed evolution methodology[29], in which randomly mutated pools are screened. A related approach is the experimental evolution[30] of cells in chemostats[31] or by serial dilution[32,33]. The advantage of these methods is that more extensive and unbiased evolutionary changes can be explored, although they do not directly reveal why trajectories are chosen. Together, these developments may change the character of molecular evolution research from one that is primarily sequence-based to one that explicitly incorporates structure, function and fitness.

1. Darwin, C. *On the Origin of Species by Means of Natural Selection* Ch VI (Murray, London, 1859).
2. Pauling, L. & Zuckerkandl, E. Chemical paleogenetics; Molecular "restoration studies" of extinct forms of life. *Acta Chem. Scand. A* **17,** S9–S16 (1963).
3. Maynard Smith, J. Natural selection and the concept of a protein space. *Nature* **225,** 563–564 (1970).
4. Malcolm, B. A. *et al.* Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. *Nature* **345,** 86–89 (1990).
5. Stackhouse, J., Presnell, S. R., McGeehan, G. M., Nambiar, K. P. & Benner, S. A. The ribonuclease from an extinct bovid ruminant. *FEBS Lett.* **262,** 104–106 (1990).
6. Ugalde, J. A., Chang, B. S. W. & Matz, M. V. Evolution of coral pigments recreated. *Science* **305,** 1433 (2004).
7. Thornton, J. W. Resurrecting ancient genes: Experimental analysis of extinct molecules. *Nature Rev. Genet.* **5,** 366–375 (2004).
8. Wright, S. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proc 6th Int. Cong. Genet.* **1,** 356–366 (1932).
9. Gillespie, J. H. *The Causes of Molecular Evolution* (Oxford Univ. Press, Oxford, 1991).
10. Kauffman, S. A. *The Origins of Order: Self-organization and Selection in Evolution* (Oxford Univ. Press, Oxford, 1993).
11. Gavrilets, S. *Fitness Landscapes and the Origin of Species* (Princeton Univ. Press, Princeton, 2004).
12. van Nimwegen, E. & Crutchfield, J. P. Metastable evolutionary dynamics: crossing fitness barriers or escaping via neutral paths? *Bull. Math. Biol.* **62,** 799–848 (2000).
13. Weinreich, D. M., Watson, R. A. & Chao, L. Sign epistasis and genetic constraint on evolutionary trajectories. *Evol. Int. J. Org. Evol.* **59,** 1165–1174 (2005).
14. Weinreich, D. M., Delaney, N. F., DePristo, M. A. & Hartl, D. L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312,** 111–114 (2006).
15. Kimura, M. On the probability of fixation of mutant genes in a population. *Genetics* **47,** 713–719 (1962).
16. Poelwijk, F. J., Kiviet, D. J. & Tans, S. J. Evolutionary potential of a duplicated repressor-operator pair: simulating pathways using mutation data. *PLoS Comput. Biol.* **2,** e58 (2006).
17. DePristo, M. A., Weinreich, D. M. & Hartl, D. L. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nature Rev. Genet.* **6,** 678–687 (2005).
18. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. *Proc. Natl Acad. Sci. USA* **103,** 5869–5874 (2006).
19. Lunzer, M., Miller, S. P., Felsheim, R. & Dean, A. M. The biochemical architecture of an ancient adaptive landscape. *Science* **310,** 499–501 (2005).
20. Zhu, G., Golding, G. B. & Dean, A. M. The selective cause of an ancient adaptation. *Science* **307,** 1279–1282 (2005).
21. Bridgham, J. T., Carroll, S. M. & Thornton, J. W. Evolution of hormone-receptor complexity by molecular exploitation. *Science* **312,** 97–101 (2006).
22. Lehming, N., Sartorius, J., Kisters-Woike, B., von Wilcken-Bergmann, B. & Müller-Hill, B. Mutant lac repressors with new specificities hint at rules for protein-DNA recognition. *EMBO J.* **9,** 615–621 (1990).
23. Barkai, N. & Leibler, S. Robustness in simple biochemical networks. *Nature* **387,** 913–917 (1997).
24. Kirschner, M. & Gerhart, J. Evolvability. *Proc. Natl Acad. Sci. USA* **95,** 8420–8427 (1998).
25. Kitano, H. Biological robustness. *Nature Rev. Genet.* **5,** 826–837 (2004).
26. Stelling, J., Sauer, U., Szallasi, Z., Doyle, F. J. III & Doyle, J. Robustness of cellular functions. *Cell* **118,** 675–685 (2004).
27. Thattai, M. & van Oudenaarden, A. Stochastic gene expression in fluctuating environments. *Genetics* **167,** 523–530 (2004).
28. Kussell, E. & Leibler, S. Phenotypic diversity, population growth, and information in fluctuating environments. *Science* **309,** 2075–2078 (2005).
29. Arnold, F. H., Wintrode, P. C., Miyazaki, K. & Gershenson, A. How enzymes adapt: lessons from directed evolution. *Trends Biochem. Sci.* **26,** 100–106 (2001).
30. Elena, S. F. & Lenski, R. E. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nature Rev. Genet.* **4,** 457–469 (2003).
31. Couñago, R., Chen, S. & Shamoo, Y. *In vivo* molecular evolution reveals biophysical origins of organismal fitness. *Mol. Cell* **22,** 441–449 (2006).
32. Lenski, R. E. & Travisano, M. Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proc. Natl Acad. Sci. USA* **91,** 6808–6814 (1994).
33. Dekel, E. & Alon, U. Optimality and evolutionary tuning of the expression level of a protein. *Nature* **436,** 588–592 (2005).
34. Hurley, J. H., Dean, A. M., Koshland, D. E. Jr & Stroud, R. M. Catalytic mechanism of NADP(+)-dependent isocitrate dehydrogenase: implications from the structures of magnesium-isocitrate and NADP+ complexes. *Biochemistry* **30,** 8671–8678 (1991).
35. Hurley, J. H., Chen, R. & Dean, A. M. Determinants of cofactor specificity in isocitrate dehydrogenase: structure of an engineered NADP+ → NAD+ specificity-reversal mutant. *Biochemistry* **35,** 5670–5678 (1996).
36. Kalodimos, C. G. *et al.* Plasticity in protein-DNA recognition: lac repressor interacts with its natural operator 01 through alternative conformations of its DNA-binding domain. *EMBO J.* **21,** 2866–2876 (2002).
37. Kopke Salinas, R. *et al.* Altered specificity in DNA binding by the lac repressor: a mutant lac headpiece that mimics the gal repressor. *ChemBioChem* **6,** 1628–1637 (2005).
38. Koradi, R., Billeter, M. & Wüthrich, K. MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14,** 51–55 (1996).

Acknowledgements We thank A. Dean, D. Hartl, J. Thornton and W. Vos for critical reading of the manuscript, and S. Tănase-Nicola for discussions. We thank A. Bonvin and R. Salinas for supplying the data for Fig. 2b. This work is part of the research programme of the Stichting voor Fundamenteel Onderzoek der Materie (FOM), which is financially supported by the Nederlandse Organisatie voor Wetenschappelijke Onderzoek (NWO).

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence should be addressed to S.J.T. (tans@amolf.nl).

ARTICLES

Coupling substrate and ion binding to extracellular gate of a sodium-dependent aspartate transporter

Olga Boudker[1]*†, Renae M. Ryan[1]*†, Dinesh Yernool[1]†, Keiko Shimamoto[3] & Eric Gouaux[1,2]†

Secondary transporters are integral membrane proteins that catalyse the movement of substrate molecules across the lipid bilayer by coupling substrate transport to one or more ion gradients, thereby providing a mechanism for the concentrative uptake of substrates. Here we describe crystallographic and thermodynamic studies of Glt_{Ph}, a sodium (Na^+)-coupled aspartate transporter, defining sites for aspartate, two sodium ions and D,L-threo-β-benzyloxyaspartate, an inhibitor. We further show that helical hairpin 2 is the extracellular gate that controls access of substrate and ions to the internal binding sites. At least two sodium ions bind in close proximity to the substrate and these sodium-binding sites, together with the sodium-binding sites in another sodium-coupled transporter, LeuT, define an unwound α-helix as the central element of the ion-binding motif, a motif well suited to the binding of sodium and to participation in conformational changes that accompany ion binding and unbinding during the transport cycle.

Life depends on the coordinated movement of molecules and ions across the membranes of cells and organelles, catalysed by specific membrane proteins—channels and pumps. Channels create continuous transmembrane pores and facilitate the movement of ions and electrolytes down their electrochemical gradients[1]. In contrast, pumps couple the thermodynamically unfavourable movement of substrates to energy stored in the form of ATP or electrochemical gradients[2,3]. Secondary transporters are 'pumps' that couple movements of substrates and one or more ions and thus 'push' their substrates up steep thermodynamic hills by harnessing pre-existing ion gradients[4]. Sodium-driven secondary transporters are particularly important in biology, catalysing the uptake of neurotransmitters[5], sugars[6] and amino acids[4,7]. The mechanisms by which these transporters couple substrate and sodium transport are currently fundamental unanswered questions in biology.

Mammalian excitatory amino acid transporters (EAATs) catalyse the uptake of the neurotransmitter glutamate from chemical synapses[8] and are representatives of a large family of secondary transporters that move acidic and neutral amino acids, as well as dicarboxylic acids, across the membranes of prokaryotic and eukaryotic cells[9]. By coupling the neurotransmitter uptake to the co-transport of three sodium ions and one proton, and the countertransport of a potassium (K^+) ion, EAATs pump substrates against concentration gradients of up to several thousandfold[10–12]. The recently determined structure of Glt_{Ph}, an EAAT homologue from *Pyroccocus horikoshii*, reveals a trimeric bowl-shaped architecture with an aqueous basin facing the extracellular solution and reaching halfway across the membrane bilayer (Supplementary Fig. S1a)[13]. Each protomer harbours eight transmembrane segments, two re-entrant helical hairpins, and independent substrate translocation pathways[14,15]. The first six transmembrane segments form a distorted 'amino-terminal cylinder' and provide all interprotomer contacts, whereas transmembrane segments TM7 and TM8, together with hairpins HP1 and HP2, coalesce to form a highly conserved core within the amino-terminal cylinder.

Nestled between the tips of HP1 and HP2, within a site secluded from aqueous solution, is non-protein electron density that we have previously suggested to represent a bound substrate, the identity of which was not known at the time of the initial structure determination[13]. The location of this buried substrate site is reminiscent of the occluded location of the leucine site in LeuT[16], which is unrelated in amino acid sequence[5]. The structures of both Glt_{Ph} and LeuT indicate that the reaction cycle of these secondary transporters might involve at least three states: open to the outside, occluded, and open to the inside. Here we determine the conformational transitions of the Glt_{Ph} transporter that allow substrates and ions to reach their binding sites from the extracellular solution, explain the mechanism that couples the opening and closing of the extracellular 'gate' to substrate, ion and inhibitor binding, and show that Glt_{Ph} and LeuT share a common sodium-ion-binding motif.

Substrate and ion specificity

Glt_{Ph} from *Pyrococcus horikoshii* shares as much as 36% amino acid sequence identity with the eukaryotic EAATs and exhibits much greater conservation in regions of functional importance[17–23]. Radiolabelled flux experiments revealed that Glt_{Ph} preferentially catalyses the sodium-dependent uptake of ^{3}H-aspartate over ^{3}H-glutamate (Fig. 1a, and Supplementary Fig. S1b). Sodium was overwhelmingly the most effective at driving transport, with lithium supporting very slow uptake (Supplementary Fig. S1b) and potassium yielding no measurable transport activity (data not shown). Aspartate uptake was only modestly stimulated by conditions of an internal negative

[1]Department of Biochemistry and Molecular Biophysics, and [2]Howard Hughes Medical Institute, Columbia University, 650 West 168th Street New York, New York 10032, USA. [3]Suntory Institute for Bioorganic Research, Wakayamadai, Shimamoto-cho, Misima-gun, Osaka 618-8503, Japan. †Present addresses: Department of Physiology and Biophysics, Weill Medical College of Cornell University, 1300 York Avenue, New York, New York 10021, USA (O.B.); National Institute of Neurological Disorders and Stroke, National Institutes of Health, 35 Convent Drive, Bethesda, Maryland 20892, USA (R.M.R.); Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907, USA (D.Y.); Vollum Institute and Howard Hughes Medical Institute, Oregon Health and Science University, Portland, Oregon 97239, USA (E.G.).
*These authors contributed equally to this work.

potential (Fig. 1a); therefore, although indicative, this does not provide conclusive evidence of electrogenic transport. Strikingly, D,L-threo-β-benzyloxyaspartate (TBOA; Supplementary Fig. S1c), a competitive inhibitor of eukaryotic glutamate transporters[24], inhibited aspartate uptake with a 50% inhibitory concentration (IC_{50}) of 3.3 ± 1.5 μM (mean ± s.d.) (Fig. 1a, and Supplementary Fig. S1d).

To measure substrate and ion binding to Glt_{Ph}, we developed a fluorescence-based assay by introducing a single tryptophan residue (L130W; Supplementary Fig. S1a). Upon titration of Glt_{Ph}-L130W with L-aspartate, we observed a substantial increase in intrinsic protein fluorescence; from these data we calculated dissociation constant (K_d) values for L-aspartate and D-aspartate of about 1 nM and 10 nM in 200 mM NaCl, respectively (Fig. 1b). In agreement with the uptake experiments, titration of Glt_{Ph}-L130W with aspartate in the presence of potassium, choline and ammonium did not yield significant changes in tryptophan fluorescence. Lithium (Li^+) supported aspartate binding, although the K_d measured in the presence of 200 mM LiCl was about 1,000-fold higher than that measured in 200 mM NaCl. In contrast to aspartate, glutamate bound only weakly to Glt_{Ph} (Fig. 1c); this bacterial homologue therefore differs from the mammalian EAATs, which transport L-glutamate, L-aspartate and D-aspartate with apparent micromolar affinities[25–27].

To define the position of aspartate in the substrate-binding site we exploited the fact that L-cysteine sulphinic acid (L-CS; Fig. 1c and Supplementary Fig. S1c) binds tightly to Glt_{Ph} and harbours an anomalous scatterer (sulphur), thereby allowing us to position the sulphinic moiety in anomalous difference Fourier maps derived from cocrystals of Glt_{Ph} and L-CS. Together with simulated annealing omit maps calculated from aspartate-bound crystals, we fitted aspartate into its sausage-shaped electron density by assuming that the β-carboxylate of aspartate occupied the same position as the sulphinic acid group of L-CS (Supplementary Fig. S2b, c and Supplementary Table S1).

Aspartate is completely buried within a polar chamber located halfway across the membrane bilayer and formed by the tips of HP1 and HP2, the unwound region of TM7 (NMDGT motif) and polar residues of amphipathic TM8, regions previously implicated in substrate binding and translocation[28]. Key interactions involve the α-amino and carboxylate groups of aspartate and R276 (HP1), V355 (HP2) and D394/N401 (TM8) as well as the β-carboxylate and T314 (TM7), G359 (HP2) and R397 (TM8). Residues involved in aspartate coordination (Fig. 1d) are mostly conserved, with notable substitutions of D394, which is close to the α-amino group and is a serine residue in dicarboxylate transporters[29,30] and R397, which is proximal to the β-carboxylate and is a neutral amino acid or an aspartate in neutral amino acid transporters of eukaryotes[31,32] and bacteria[9,33], respectively.

Extracellular gate

The occluded substrate-binding site observed in Glt_{Ph} raises the question of how aspartate reaches this site from the extracellular or intracellular solution, and what portions of the transporter act as gates, controlling access to the binding site. A clue to the identity of the extracellular gate came from solving the crystal structure of Glt_{Ph} in complex with the non-transportable blocker TBOA (Supplementary Fig. S1c). The overall structure of the transporter is similar in the

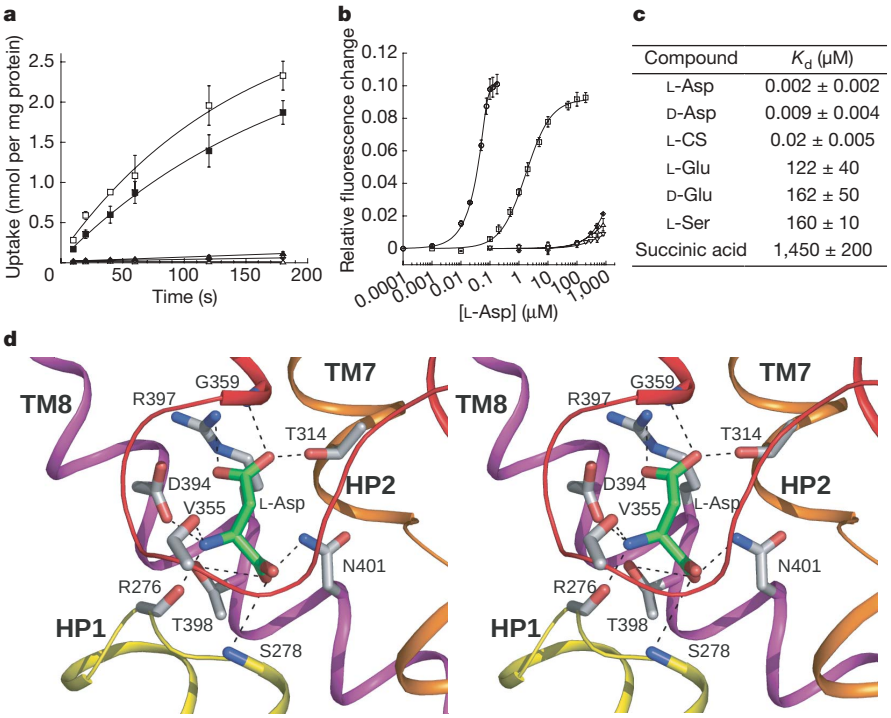


Figure 1 | Glt_{Ph} is an aspartate-specific sodium-driven transporter. **a**, Ion specificity and inhibition of ^{3}H L-aspartate uptake by Glt_{Ph}. Loaded proteoliposomes were diluted into solutions containing the following ions (in mM; see Methods; initial rates are in parentheses): 100 Na^+, 200 K^+ (filled squares, 14 pmol mg s^{-1}); 100 Na^+, 200 choline, internal negative (open squares, 20 pmol mg s^{-1}); 100 Li^+, 200 K^+ (diamonds, 0.1 pmol mg s^{-1}); 100 Na^+, 200 K^+, 1 mM TBOA (filled triangles, 0.5 pmol mg s^{-1}); and 100 choline, 200 K^+ (open triangles, −0.04 pmol mg s^{-1}). **b**, Fluorescence changes observed in the Glt_{Ph}-L130W mutant on titration with L-aspartate in the presence of the chloride salts of the following ions (each at 200 mM): Na^+ (circles), Li^+ (squares), K^+ (diamonds), NH_4^+ (triangles) and choline (inverted triangles). Uptake experiments with the L130W mutant show that it is active, although the level of activity is about 15-fold lower than for the wild-type transporter (Supplementary Fig. S1a). **c**, Dissociation constants for potential substrates in 200 mM NaCl. **d**, Stereo view of the aspartate-binding site showing HP1 (yellow), TM7 (orange), HP2 (red) and TM8 (magenta). A remarkable number of polar contacts solvate the highly charged substrate and include interactions with D394, main-chain carbonyls of R276 (HP1) and V355 (HP2), the amide nitrogen of N401 (TM8), the hydroxyl of T398 (TM8), the main-chain nitrogen of S278, the guanidinium group of R397 (TM8), the hydroxyl of T314 (TM7) and the main-chain nitrogen of G359 (HP2). Results in **a**–**c** are means ± s.d.

aspartate-bound and TBOA-bound complexes with the important exception of two regions: in the TBOA-bound structure HP2 adopts an 'open' conformation, moving as much as 10 Å from its position in the aspartate-bound complex towards the 3L4 loop, with the 3L4 loop also shifting closer to HP2, enabling direct contacts (Fig. 2a, and Supplementary Fig. S3b).

Guided by anomalous difference maps calculated from diffraction data of the complex with 3-bromo-TBOA (3-Br-TBOA), we modelled TBOA into the excess electron density observed in Glt_{Ph}–TBOA density maps (Fig. 2a, and Supplementary Fig. S3a). We found that the aspartate moiety of TBOA binds in the substrate-binding site, in harmony with the observation that TBOA has competitive inhibition kinetics[24] and that the bulky benzyl group lodges against the tip of HP2, propping it in an open conformation. The benzyl group contacts HP2 near the main-chain atoms of G359 and interacts with M311 in TM7. The conservation of M311 provides a structural explanation for the broad yet potent inhibition by TBOA of eukaryotic glutamate transporters and of the evolutionarily distant Glt_{Ph}.

The movement of HP2 in the TBOA-bound complex exposes the substrate-binding site to the extracellular solution (Fig. 2a, c) and raises the question whether HP2 adopts an 'open' conformation before aspartate binding when the transporter is in the 'apo' state. We prepared substrate-depleted Glt_{Ph} crystals (see Methods) and examined threefold averaged electron density maps corresponding to the HP2 region. We found that HP2 predominantly occupies an 'open' conformation, essentially indistinguishable from that seen in the TBOA-bound state (Fig. 2b). In unaveraged maps, however, subunit C and to a smaller degree subunit A showed density for HP2 in a closed conformation, perhaps because of residual aspartate, lattice interactions or stochastic fluctuation of HP2. Nevertheless, these results indicate that in the absence of substrate, HP2 can adopt an 'open' conformation, rendering the substrate-binding site accessible to extracellular solution (Fig. 2c).

Closure of HP2, as seen in the aspartate-bound state, creates a crevice between HP2, the 3L4 loop and TM4a (Fig. 2d). In all three subunits an elongated electron density feature is observed in this crevice, extending from the solvent-filled basin to the outer, lipid-exposed 'walls' of the transporter and lying nearly parallel to the membrane plane (Supplementary Fig. S3c–e); we have modelled this feature as the alkyl chain of a lipid molecule. Identification of a state-dependent lipid binding pocket in Glt_{Ph} indicates that similar pockets might exist in eukaryotic transporters, perhaps providing sites for lipid binding and the modulation of transporter activity[34].

Sodium coupling

In a manner similar to substrate transport by the EAATs[35], transport of aspartate by Glt_{Ph} is sodium dependent. Fluorescence binding assays with Glt_{Ph}-L130W revealed that aspartate binding itself was sodium dependent. At sodium concentrations above 50 mM, the K_d values for aspartate were lower than the protein concentrations used in the assays, resulting in binding isotherms that were anomalously steep (see Supplementary Information). At lower sodium concentra-

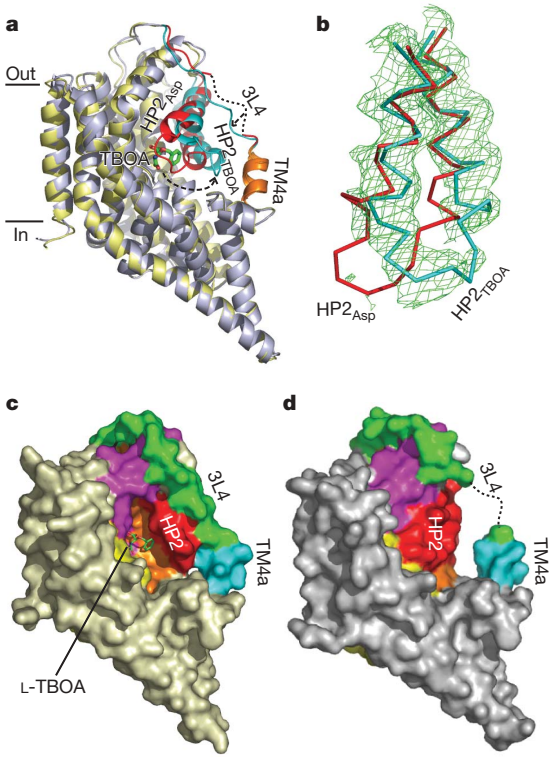


Figure 2 | HP2 is the extracellular gate. **a**, Overlay of Glt_{Ph}–L-Asp (grey) and Glt_{Ph}–TBOA (gold) complexes. Single protomers are shown in a schematic representation and the regions undergoing significant conformational changes are in red (L-aspartate complex) and in cyan (TBOA complex). TM4a is orange. The model of TBOA is shown in stick representation. **b**, Simulated annealing $2F_o - F_c$ omit electron density map for HP2 in apo Glt_{Ph} (green mesh; 1σ). α-Carbon traces for HP2 in the L-aspartate (red) and TBOA (cyan) complexes are shown. **c**, **d**, Solvent-accessible surface view of a single subunit in complex with TBOA (**c**) and with aspartate (**d**); the 3L4 loop, TM7, HP1, HP2 and TM8 are coloured as in Supplementary Fig. S1a. In **d** the 'gap' between HP2 and 3L4/TM4a is partly filled with lipid (see also Supplementary Fig. S3c–e).

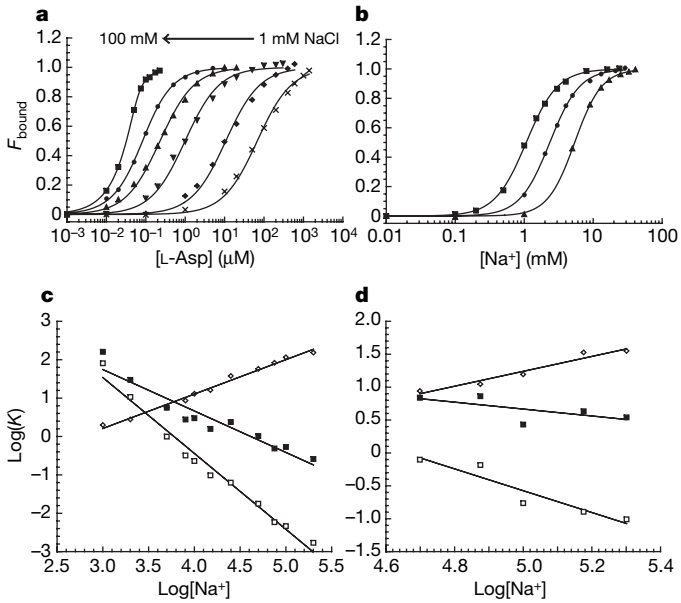


Figure 3 | Binding of ligands and sodium to Glt_{Ph}. **a**, **b**, Sodium dependence of L-aspartate binding (**a**) and L-aspartate dependence of sodium binding (**b**; triangles, 1 μM; circles, 10 μM; squares, 100 μM L-Asp). The fraction of bound transporter was calculated by dividing the relative fluorescence change of Glt_{Ph}-L130W on the addition of L-aspartate or sodium by the total change at the end of the titration. Endpoint sodium concentrations are shown above the graph. Solid lines through the data are fits to the equations described in the Supplementary Information. **c**, Logarithmic plots of L-aspartate K_d values (open squares; slope = -2.0 ± 0.1) and TBOA K_i values (filled squares; slope = -1.1 ± 0.1) against log([Na^+](μM)) are shown for Glt_{Ph}-L130W, with data from the fluorescence assay. Differences between log K_d values for L-aspartate and log K_i values for TBOA (diamonds; slope = 0.9 ± 0.03) are also plotted. **d**, Logarithmic plots of L-aspartate (open squares; slope = -1.7 ± 0.3) and sodium (filled squares; slope = -0.3 ± 0.3) concentrations from fluorescence binding studies of the Glt_{Ph}-D405N/L130W mutant. Differences between log K_d values for L-aspartate and log K_i values for TBOA (diamonds; slope = 1.1 ± 0.15) are also plotted.

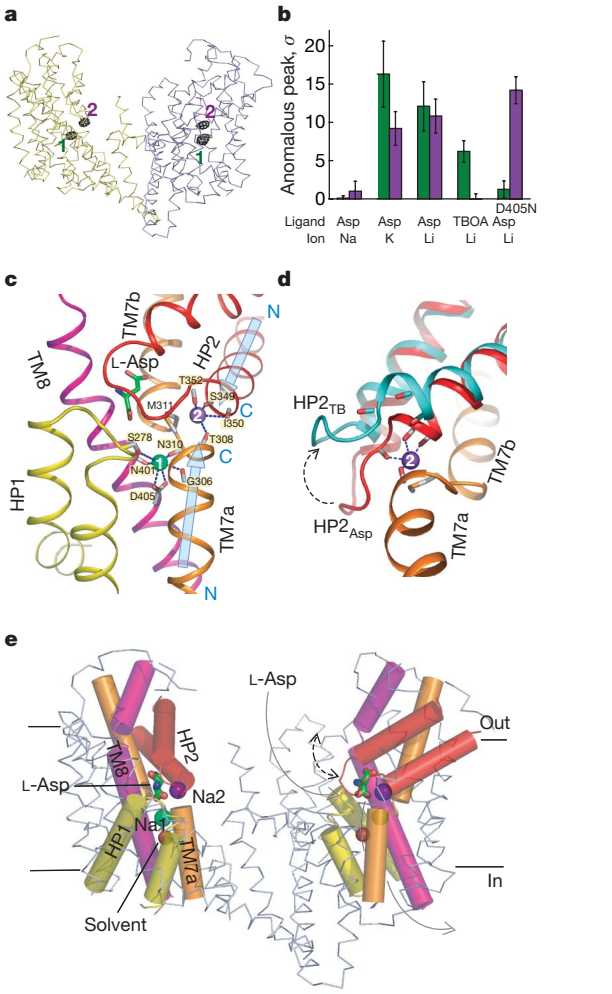


Figure 4 | Sodium-binding sites in Glt_{Ph}. **a**, Two subunits of Glt_{Ph} are shown. Anomalous difference Fourier map calculated for Glt_{Ph}–L-aspartate complex crystals soaked in 50 mM Tl^+ and 100 mM Li^+ and contoured at 5.5σ (black mesh). Two ion-binding sites are designated 1 (green) and 2 (purple). **b**, Heights of peaks on anomalous difference maps for crystals soaked in 50 mM Tl^+ and 100 mM Na^+, K^+ or Li^+. Peaks are mean values derived from the three protomers and one (Na^+) or two (K^+, Li^+) data sets. Anomalous peaks for TBOA-bound Glt_{Ph} and aspartate-bound D405N soaked in Tl^+ and Li^+ are also shown. When no clear peaks were detected on the map, values of the electron density at sites 1 or 2 were employed. The green bars represent the peak heights at site 1 and the purple bars the peak heights at site 2. Results are means ± s.d. **c**, Oxygen atoms that are within 3.5 Å of the sodium ions are labelled and connected to the sodium ions by dashed lines. Light blue arrows represent the dipole moments of helices TM7a and HP2a. **d**, Opening of HP2 observed in the TBOA-bound structure destroys sodium site 2. The dashed arrow indicates the direction of HP2 motion in the TBOA-bound state. Sodium bound at site 2 in aspartate-bound Glt_{Ph} is shown as a sphere, with coordinating oxygen interactions depicted by dashed lines. **e**, Location of sodium-binding sites on the permeation pathway of aspartate. Two protomers of Glt_{Ph} are shown. N-terminal cylinders are in ribbon representation, the TM helices in the C-terminal protein cores are shown as cylinders and bound aspartate is shown in stick representation. HP2 (red) serves as an extracellular gate and opens to afford aspartate access to the binding site. Sodium 2 (purple) serves as a lock on the gate, providing additional energy necessary for its closure. Below the substrate-binding site is sodium 1 (green) and bound solvent (red sphere). The proposed intracellular gate is formed by HP1 (yellow), TM7 (orange) and TM8 (magenta), which are held together by sodium 1. The proposed permeation pathway for the substrate is shown as a grey line and motions of HP2 are shown as a dashed double-headed arrow.

tions, however, aspartate bound less tightly and the isotherms were well fitted with the Hill equation with coefficients near unity, suggesting that aspartate binds independently to each protomer (Fig. 3a)[14]. Measurements of aspartate binding to wild-type Glt_{Ph} by isothermal titration calorimetry (ITC) yielded similar results, reinforcing the conclusion that aspartate binding is not cooperative (Supplementary Fig. S4a).

Sodium titrations of Glt_{Ph}-L130W in the presence of aspartate also resulted in fluorescence changes, and yielded apparent K_d values for sodium in the millimolar range and Hill coefficients of about 2, suggesting the coupled binding of at least two sodium ions (Fig. 3b). A plot of log aspartate K_d against log sodium concentration was well fitted by a straight line with a slope of about 2, which corresponds to the number of sodium ions coupled to the binding of an aspartate molecule (Fig. 3c, Supplementary Fig. S4c and Supplementary Information). Measurements from ITC experiments yielded similar K_d values and slopes between 1.6 and 2.0, providing further support for the conclusion that each subunit independently binds one aspartate and at least two sodium ions.

The sodium dependence of TBOA binding, by contrast, was weaker, and plotting the apparent inhibition constants (K_i) for TBOA as a function of sodium concentration yielded linear fits with slopes of about 1.1 (Fig. 3c). Here, competition experiments with aspartate were performed because titration of TBOA with unliganded Glt_{Ph}-L130W did not yield a fluorescence signal. In ITC experiments with wild-type Glt_{Ph} (Supplementary Fig. S4b, c) we also found that TBOA binding was coupled to about one sodium ion. The ITC experiments yielded K_d values that were severalfold lower than the K_i values from the fluorescence experiments, and this might have been due to unfavourable interactions between HP2 in the open state and the tryptophan residue in L130W-Glt_{Ph} (Supplementary Fig. S3b). Nevertheless, these two independent approaches show that TBOA binding is coupled to about one sodium ion, whereas aspartate binding is coupled to about two.

Sodium-binding sites

To define the positions of the sodium sites in Glt_{Ph} we employed thallium(I) (Tl^+), a monovalent ion with a robust anomalous scattering signal. Examination of anomalous difference Fourier maps of crystals soaked in thallous nitrate revealed two strong peaks per subunit (Fig. 4a), resulting from the partial occupancy (about 0.2–0.4) of Tl^+ at these sites. To evaluate whether these sites, defined as sites 1 and 2, were specific for sodium, competition experiments were performed by soaking crystals of the Glt_{Ph}–aspartate complex in solutions containing Tl^+ and either Li^+, Na^+ or K^+ (Fig. 4b). Only sodium diminished the thallium anomalous density peaks, supporting the contention that the sites labelled by thallium are genuine sodium sites. We suggest that lithium did not compete with Tl^+ simply because the binding of Li^+ is about 10–30-fold weaker than sodium, as estimated from the ~1,000-fold weaker binding of aspartate in the presence of Li^+ (Fig. 1b). Similar crystallographic studies of the Glt_{Ph}–TBOA complex revealed only one strong Tl^+ peak in anomalous difference Fourier maps (Fig. 4b), located at site 1.

The sodium sites are near the bound aspartate, although neither is in direct contact (Fig. 4c). Sodium site 1 is buried deeply within the protein, below aspartate, and is coordinated by three carbonyl oxygens in TM7 and TM8, a carboxyl group of D405 in TM8 and possibly by a hydroxyl oxygen of S278 in HP1. Sodium 2 is buried just under HP2 and seems to have only four coordinating carbonyl oxygens located in TM7 and HP2. The carboxy termini of helices TM7a and HP2a point towards site 2, indicating that their dipole moments might stabilize the bound ion (Fig. 4c). Involvement of HP2 in the formation of site 2 explains why this site is absent in the TBOA-bound state: HP2 is no longer proximal to TM7a and therefore cannot contribute coordinating oxygen atoms (Fig. 4d).

Why does Tl^+ label the sodium sites in Glt_{Ph}? The Pauling radius of Tl^+ ($r = 1.40$ Å) is most similar to that of K^+ ($r = 1.33$ Å) and

substantially larger than that of Na^+ ($r = 0.95$ Å), and Tl^+ is typically employed as a heavy-ion analogue of potassium, although even in complexes with crown ethers Tl^+ and K^+ often form completely different structures[36]. Tl^+ is a highly polarizable ion, and we suggest that it forms favourable interactions with the sulphur of M311, which is sandwiched between the sodium-binding sites, 3.3–5.0 Å from the sodium ion positions (Fig. 4c). Thus, in addition to the oxygen ligands that define the smaller sodium-selective binding sites, Tl^+, but not K^+, is able to exploit the presence of M311 and bind to these sodium-selective sites. Although Tl^+ did not support ^{3}H-aspartate uptake in flux experiments, Tl^+ can replace sodium in the binding of aspartate to Glt_{Ph}, in a similar manner to Li^+, and Tl^+ reversibly inhibited Na^+-driven aspartate transport (Supplementary Fig. S5).

D405N mutant

The carboxylate group of D405 coordinates sodium 1 (Fig. 4c) and, whereas it is conserved in mammalian orthologues, it is an asparagine residue in some bacterial proton-dependent transporters[9]. To probe the role of D405 in sodium binding and in its coupling to aspartate binding, we made the D405N mutant and performed crystallographic and binding studies. The crystal structure of the Glt_{Ph}-D405N–aspartate complex was indistinguishable from that of the wild-type Glt_{Ph}–aspartate complex. However, on inspection of anomalous Fourier difference maps calculated from data measured on crystals soaked in Tl^+, we found a strong peak at site 2 but no peak at site 1 (Fig. 4b), suggesting that the mutation weakened the binding of Tl^+, and by analogy that of Na^+, to site 1. The D405N mutation diminished aspartate binding about 100-fold, resulting in a K_d of about 100 nM in 200 mM NaCl (data not shown).

Substrate and inhibitor binding was still sodium dependent, and logarithmic plots of aspartate K_d and TBOA K_i values against sodium concentration gave straight lines with slopes of 1.7 and 0.5, respectively (Fig. 3d). In agreement with the loss of Tl^+ binding to site 1, the sodium dependence for aspartate and TBOA binding to the D405N mutant was decreased. Because the apparent couplings were decreased by 0.3–0.5 rather than by about 1.0, we suggest that the binding of sodium to site 1 is only partly coupled to aspartate and TBOA binding and that there may be a third sodium site, one that is resistant to Tl^+ substitution. A recent study on the EAAC1 glutamate transporter indicates that the mutation equivalent to D405N in eukaryotic transporters does not abrogate sodium binding to the glutamate-free state[37] and is also consistent with one or more additional sodium binding sites.

TM7 harbours a sodium-binding motif

We compared the sodium sites in Glt_{Ph}, together with key elements of the local protein structure, with the sodium sites identified in LeuT (PDB code 2A65). Strikingly, key unwound transmembrane segments of Glt_{Ph} (TM7) and LeuT (TM1)[16] reveal similarity in their local protein conformations and in the relative disposition of sodium-binding sites (Fig. 5a–c). In Glt_{Ph} and LeuT, sodium site 1 is defined, in part, by coordinating carbonyl oxygen atoms occupying nearly equivalent positions, and the ion in site 2 is also coordinated by an equivalently positioned carbonyl oxygen atom. Even though Glt_{Ph} and LeuT are unrelated in amino acid sequence and three-dimensional structure, elements of protein structure surrounding the sodium sites are somewhat similarly organized (Fig. 5d, e). For example, TM8 in both Glt_{Ph} and LeuT fits into a groove created by the unwound portions of TM7 and TM1, respectively, and in so doing it provides key coordinating oxygen atoms to sodium site 1. Elaboration of sodium site 2, in turn, is provided by exposed carbonyl oxygen atoms at the C termini of two roughly equivalently positioned helices, which are HP2 in Glt_{Ph} and TM6 in LeuT.

Taken together, the similarities in local protein structure and coordination of the sodium sites in Glt_{Ph} and LeuT indicate that sodium dependent transporters might possess a common sodium-ion-binding motif, as exemplified by TM7 and TM1 in Glt_{Ph} and LeuT. This motif has the following features (Fig. 5c): first, a break in helix structure of three to five residues that opens up the peptide chain to permit sodium ion coordination by main-chain carbonyl oxygen atoms; second, a site 1 defined by carbonyl oxygen atoms occupying the first and fourth (LeuT) or fifth (Glt_{Ph}) positions of the motif; and third, a site 2 defined by the carbonyl oxygen atom at the third position. This motif not only satisfies the requirements for sodium ion coordination but also provides a mechanism by which ion binding and unbinding can be coupled to local and long-distance conformational changes through the bending and twisting of the associated helical elements.

Mechanism

A synthesis of the structural and biophysical experiments described here, together with previous studies on glutamate transporters[38], allows us to propose a mechanism for coupling sodium and aspartate binding, movement of HP2, and inhibition of transport by TBOA

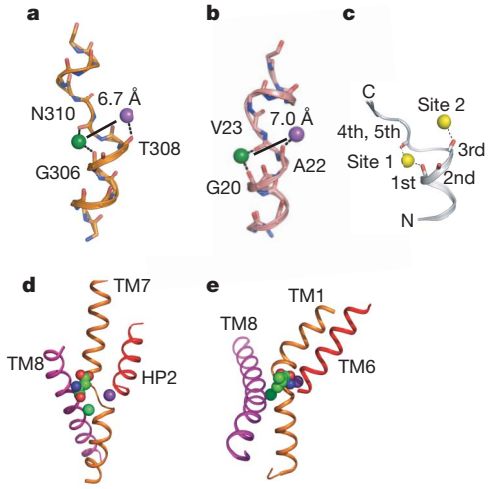


Figure 5 | Sodium-binding motif. **a**, TM7 of Glt_{Ph}, showing ions at sites 1 and 2, coloured according to Fig. 4c. **b**, TM1 of LeuT and sodium ions 1 (green) and 2 (purple). **c**, The unwound region of Glt_{Ph} TM7 defines a monovalent-ion-binding motif in sodium-dependent transporters. **d**, **e**, Substrate, sodium ions and key transmembrane regions of Glt_{Ph} (**d**) and LeuT (**e**).

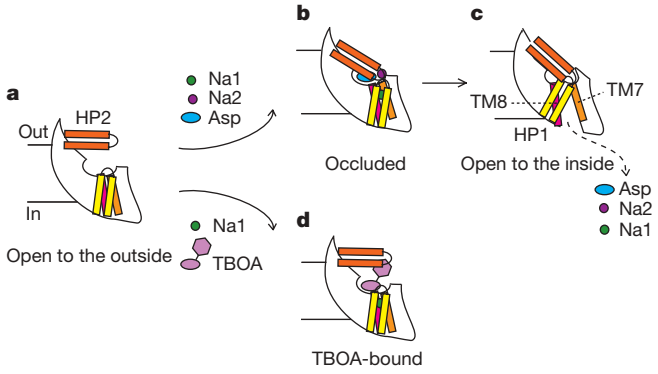


Figure 6 | Schematic mechanism illustrated by a diagram of a single subunit. **a**, Apo state with HP2 in an open conformation. **b**, Sodium ions 1 and 2 and aspartate bind and induce closure of HP2, yielding the occluded state. **c**, Opening of the internal gate allows the release of aspartate and sodium to the cytoplasm and may involve the movement of HP1, TM7 and TM8. In this state, HP2 may form additional interactions with the protein core, stabilizing HP2 in a closed conformation. **d**, TBOA binding blocks transport by stabilizing HP2 in an open conformation and precluding the binding of sodium to site 2.

(Fig. 6). Beginning with the apo state, we suggest that sodium ions bind to sites 1 and 2, together with aspartate binding to the substrate site, and that these binding events are tightly coupled to each other and induce the closure of HP2. Inhibition of transport by TBOA occurs because TBOA blocks HP2 in an open conformation, preventing sodium from binding to site 2 and halting further conformational changes along the transport cycle.

The location of the sodium sites relative to aspartate and to portions of the transporter that are either known to move during transport (HP2)[15,20,21,39–41] or that are likely to rearrange but have not yet been proved to do so, is highly suggestive (Fig. 4e). We note that site 1 is 'below' aspartate, located closer to the cytoplasm, whereas site 2 is 'above' aspartate, and nearer to the extracellular solution. Site 2 has a key function in stabilizing the extracellular gate (HP2) in a closed conformation, and we suggest that site 1 might similarly stabilize a closed conformation of the intracellular gate. Although the composition of the intracellular gate is unknown, HP1, TM7a and TM8 are likely to be important. Indeed, between sodium site 1 and the cytoplasmic surface of the transporter, in both the aspartate-bound and TBOA-bound structures, we find excess, non-protein, electron density occluded in a pocket between HP1, TM7a and TM8 (Fig. 4e), which we interpret as trapped solvent. On further conformational changes, this solvent-filled cavity may expand and provide a pathway for aspartate and sodium to reach the cytoplasm, probably along the polar face of TM8.

We speculate that there is a mechanistically symmetrical relation between sodium site 2/HP2 and sodium site 1/HP1. As we have shown that site 2 is coupled to the opening and closing of HP2, site 1 may be coupled to the opening and closing of HP1 and other components of the cytoplasmic gate (Fig. 6), thus providing a mechanism for the sodium-coupled alternating access to the aspartate-binding site and elevating sodium ions to essential 'gate-keepers' of Glt_{Ph} and other members of this sodium-coupled transporter family.

METHODS

Ligand and ion binding. Glt_{Ph} and Glt_{Ph}-L130W were expressed as His_8 fusion proteins and purified as described previously[13]. Proteins were concentrated and dialysed against HEPES/Tris buffer, pH 7.4, containing 200 mM choline chloride, 1 mM NaCl and 1 mM n-dodecyl-β-D-maltoside. For fluorescence assays Glt_{Ph}-L130W was diluted about 200-fold (final protein concentration about 100 nM) at 20 °C into a 3-ml quartz cell containing buffer with variable NaCl concentrations. Tryptophan fluorescence was excited at 295 nm and emission was measured at 334 nm. Fluorescence changes were normalized for the initial protein fluorescence, and binding isotherms were analysed as described in Supplementary Information. For ITC experiments Glt_{Ph} was diluted to 4 μM into the reaction cell of a Microcal VP-ITC calorimeter and titrations were performed at 20 °C by injecting 3–5 μl of syringe solution containing potassium salts of L-aspartic acid (100 μM) or L-TBOA (125 μM). Binding isotherms were analysed with Microcal software.

Transport assay. Liposomes were prepared with *Escherichia coli* polar lipid extract and egg phosphatidylcholine in a 3:1 (w/w) ratio. Preformed liposomes were treated with Triton X-100 at a 2:1 (w:w) lipid to detergent ratio, and Glt_{Ph} was reconstituted at a protein-to-lipid ratio of 1:330 (w/w). Transport was assayed at 30 °C (ref. 42). Typically, the uptake reaction was initiated by diluting proteoliposomes loaded with 20 mM HEPES/Tris buffer pH 7.4, 200 mM KCl and 100 mM choline chloride into buffer containing 20 mM HEPES/Tris pH 7.4, 200 mM KCl, 100 mM XCl ($X = Na^+$, Li^+, K^+ or Ch^+), 1 μM valinomycin and 75 nM L-^{3}H-aspartate or 75 nM L-^{3}H-glutamate. Background was defined as the counts observed when the proteoliposomes were diluted into the buffer with which they were loaded. To test the electrogenicity of transport, KCl in the uptake buffer was replaced by choline chloride (inside negative). TBOA dose–response measurement was performed in the presence of 75 nM L-^{3}H-aspartate and uptake was performed for 3 min. For Tl^+ experiments, liposomes were loaded with 100 mM potassium methanesulphonate, 20 mM HEPES pH 7.4, and all external salts were nitrate salts. Uptake was performed for 4 min at 30 °C. Reported values are means of at least three independent experiments and the uncertainties are the standard error of the mean.

Crystallography. The heptahistidine mutant of Glt_{Ph} (CAT7) and the CAT7-D405N mutant were expressed, purified and crystallized as described previously[13] in the presence of L-Glu, L-Asp or D-Asp or in the absence of substrates. In some cases the protein solution was supplemented with *E. coli* total lipid extract at a final concentration of about 0.1 mM before crystallization. To obtain crystals of TBOA-bound, 3-Br-TBOA-bound or L-CS-bound CAT7, protein purified in the presence of 5 mM L-Glu was supplemented with 1 mM TBOA before crystallization. Crystals were then soaked in mother liquor supplemented with 25% poly(ethylene glycol) (PEG)1000 and either 1 mM TBOA, 10 mM 3-Br-TBOA or 5 mM L-CS. In ion replacement experiments, CAT7 crystals were soaked in 25% PEG1000, 100 mM MES/Tris pH 6.0, 5 mM substrate, 2 mM n-decyl-β-D-maltoside and either 100 mM X_2SO_4 ($X = Li^+$, Na^+ or K^+) and 50 mM $TlNO_3$ or 150 mM Li_2SO4.

Diffraction data sets were indexed, integrated and scaled using the HKL-2000 package (Supplementary Table S1)[43]. Further analysis was performed with CCP4 programs[44]. Initial phases for the NAT crystal (Supplementary Table S1) were determined by rigid-body refinement in REFMAC[44] using published CAT7 coordinates (PDB code 1XFH). Phases for TB were obtained similarly, except that residues 110–130 and 337–371 were deleted from the model. Phases were further improved by rounds of manual rebuilding followed by restrained refinement in REFMAC with tight three-fold NCS restraints. A large peak on the $F_o - F_c$ and $2F_o - F_c$ difference maps in the core of the protein was modelled as a single water molecule. During the last round the restrained refinement was run with six TLS groups defined as three protomers and three substrate or TBOA molecules. Phases for all other crystals except NA and LID405N were obtained by rigid-body refinement. All electron density maps were subjected to threefold real-space averaging with RAVE software[45], except in ion competition experiments in which the heights of the anomalous peaks were determined separately in each protomer and then averaged.

Data for NA crystals, soaked in 50 mM $TlNO_3$ and 100 mM NaCl, were used to model sodium ions. There were no peaks on the anomalous difference Fourier maps, suggesting that thallium ions were replaced by sodium. In contrast, there were clear peaks on the $F_o - F_c$ and $2F_o - F_c$ maps, which were used to place sodium ions manually. The protein model with ions was further refined as described above. To estimate Tl^+ occupancies, the occupancy of the Tl^+ ions was manually adjusted, followed by threefold NCS-restrained refinement of *B*-factors, and this process was repeated several times until the *B*-factors of the Tl^+ ions were similar to those of the surrounding protein atoms.

To prepare apo crystals, CAT7 was purified and crystallized in the absence of the externally added substrate but in the presence of sodium, which is required for sustained protein stability. Under these conditions, CAT7 crystallizes in complex with a putative substrate carried over from bacterial growth medium. Repetitive soaking of crystals in solutions devoid of sodium is required to induce dissociation of the bound substrate and to produce an apo state. Although crystals prepared in this manner were of diminished quality, they were isomorphous to the aspartate complex. During data analysis HP2 was excluded from the protein model and, after rigid-body refinement, we performed real-space three-fold averaging to improve the quality of the maps, and subsequently inspected averaged and unaveraged maps.

Received 5 July; accepted 15 November 2006.
Published online 17 January 2007.

1. Hille, B. *Ion Channels of Excitable Membranes* (Sinauer Associates, Sunderland, Massachusetts, 2001).
2. Läuger, P. *Electrogenic Ion Pumps* (Sinauer Associates, Sunderland, Massachusetts, 1991).
3. Quick, M. W. (ed.) *Transmembrane Transporters* (Wiley-Liss, Hoboken, New Jersey, 2002).
4. Sobczak, I. & Lolkema, J. S. Structural and mechanistic diversity of secondary transporters. *Curr. Opin. Microbiol.* **8,** 161–167 (2005).
5. Chen, N. H., Reith, M. E. & Quick, M. W. Synaptic uptake and beyond: the sodium- and chloride-dependent neurotransmitter transporter family SLC6. *Pflugers Arch.* **447,** 519–531 (2004).
6. Wright, E. M. & Turk, E. The sodium/glucose cotransport family SLC5. *Pflugers Arch.* **447,** 510–518 (2004).
7. Wilson, T. H. & Ding, P. Z. Sodium-substrate cotransport in bacteria. *Biochim. Biophys. Acta* **1505,** 121–130 (2001).
8. Grewer, C. & Rauen, T. Electrogenic glutamate transporters in the CNS: molecular mechanism, pre-steady-state kinetics, and their impact on synaptic signaling. *J. Membr. Biol.* **203,** 1–20 (2005).
9. Slotboom, D. J., Konings, W. N. & Lolkema, J. S. Structural features of the glutamate transporter family. *Microbiol. Mol. Biol. Rev.* **63,** 293–307 (1999).
10. Kanner, B. I. & Bendahan, A. Binding order of substrates to the sodium and potassium ion coupled L-glutamatic acid transporter from rat brain. *Biochemistry* **21,** 6327–6330 (1982).
11. Zerangue, N. & Kavanaugh, M. P. Flux coupling in a neuronal glutamate transporter. *Nature* **383,** 634–637 (1996).
12. Levy, L. M., Warr, O. & Attwell, D. Stoichiometry of the glial glutamate transporter GLT-1 expressed inducibly in a Chinese hamster ovary cell line selected for low

endogenous Na^+-dependent glutamate uptake. *J. Neurosci.* **18,** 9620–9628 (1998).
13. Yernool, D., Boudker, O., Jin, Y. & Gouaux, E. Structure of a glutamate transporter homologue from *Pyrococcus horikoshii*. *Nature* **431,** 811–818 (2004).
14. Grewer, C. *et al.* Individual subunits of the glutamate transporter EAAC1 homotrimer function independently of each other. *Biochemistry* **44,** 11913–11923 (2005).
15. Koch, H. P. & Larsson, H. P. Small-scale molecular motions accomplish glutamate uptake in human glutamate transporters. *J. Neurosci.* **25,** 1730–1736 (2005).
16. Yamashita, A., Singh, S. K., Kawate, T., Jin, Y. & Gouaux, E. Crystal structure of a bacterial homologue of Na^+/Cl^--dependent neurotransmitter transporters. *Nature* **437,** 215–223 (2005).
17. Grunewald, M., Bendahan, A. & Kanner, B. I. Biotinylation of single cysteine mutants of the glutamate transporter GLT-1 from rat brain reveals its unusual topology. *Neuron* **21,** 623–632 (1998).
18. Slotboom, D. J., Lolkema, J. S. & Konings, W. N. Membrane topology of the C-terminal half of the neuronal, glial, and bacterial glutamate transporter family. *J. Biol. Chem.* **271,** 31317–31321 (1996).
19. Seal, R. P. & Amara, S. G. A reentrant loop domain in the glutamate carrier EAAT1 participates in substrate binding and translocation. *Neuron* **21,** 1487–1498 (1998).
20. Slotboom, D. J., Sobczak, I., Konings, W. N. & Lolkema, J. S. A conserved serine-rich stretch in the glutamate transporter family forms a substrate-sensitive reentrant loop. *Proc. Natl Acad. Sci. USA* **96,** 14282–14287 (1999).
21. Grunewald, M. & Kanner, B. I. The accessibility of a novel reentrant loop of the glutamate transporter GLT-1 is restricted by its substrate. *J. Biol. Chem.* **275,** 9684–9689 (2000).
22. Slotboom, D. J., Konings, W. N. & Lolkema, J. S. Cysteine-scanning mutagenesis reveals a highly amphipathic, pore-lining membrane-spanning helix in the glutamate transporter GltT. *J. Biol. Chem.* **276,** 10775–10781 (2001).
23. Grunewald, M., Menaker, D. & Kanner, B. I. Cysteine-scanning mutagenesis reveals a conformationally sensitive reentrant pore-loop in the glutamate transporter GLT-1. *J. Biol. Chem.* **277,** 26074–26080 (2002).
24. Shimamoto, K. *et al.* DL-threo-β-benzyloxyaspartate, a potent blocker of excitatory amino acid transporters. *Mol. Pharmacol.* **53,** 195–201 (1998).
25. Kanner, B. I. & Schuldiner, S. Mechanism of transport and storage of neurotransmitters. *CRC Crit. Rev. Biochem.* **22,** 1–38 (1987).
26. Nicholls, D. & Attwell, D. The release and uptake of excitatory amino acids. *Trends Pharmacol. Sci.* **11,** 462–468 (1990).
27. Arriza, J. L. *et al.* Functional comparisons of three glutamate transporter subtypes cloned from human motor cortex. *J. Neurosci.* **14,** 5559–5569 (1994).
28. Slotboom, D. J., Konings, W. N. & Lolkema, J. S. Glutamate transporters combine transporter- and channel-like features. *Trends Biochem. Sci.* **26,** 534–539 (2001).
29. Engelke, T., Jording, D., Kapp, D. & Pühler, A. Identification and sequence analysis of the *Rhizobium meliloti dctA* gene encoding the C_4-dicarboxylate carrier. *J. Bacteriol.* **171,** 5551–5560 (1989).
30. Yurgel, S. N. & Kahn, M. L. *Sinorhizobium meliloti dctA* mutants with partial ability to transport dicarboxylic acids. *J. Bacteriol.* **187,** 1161–1172 (2005).
31. Shafqat, S. *et al.* Cloning and expression of a novel Na^+-dependent neutral amino acid transporter structurally related to mammalian Na^+/glutamate cotransporters. *J. Biol. Chem.* **268,** 15351–15355 (1993).
32. Arriza, J. L. *et al.* Cloning and expression of a human neutral amino acid transporter with structural similarity to the glutamate transporter family. *J. Biol. Chem.* **268,** 15329–15332 (1993).
33. Ogawa, W., Kim, Y.-M., Mizushima, T. & Tsuchiya, T. Cloning and expression of the gene for the Na^+-coupled serine transporter from *Escherichia coli* and characteristics of the transporter. *J. Bacteriol.* **180,** 6749–6752 (1998).
34. Zerangue, N., Arriza, J. L., Amara, S. G. & Kavanaugh, M. P. Differential modulation of human glutamate transporter subtypes by arachidonic acid. *J. Biol. Chem.* **270,** 6433–6435 (1995).
35. Kanner, B. I. & Sharon, I. Active transport of L-glutamate by membrane vesicles isolated from rat brain. *Biochemistry* **17,** 3949–3953 (1978).
36. Mudring, A.-V. & Rieger, F. Lone pair effect in thallium(I) macrocyclic compounds. *Inorg. Chem.* **44,** 6240–6243 (2005).
37. Tao, Z., Zhang, Z. & Grewer, C. Neutralization of the aspartic acid residue Asp-367, but not Asp-454, inhibits binding of Na^+ to the glutamate-free form and cycling of the glutamate carrier EAAC1. *J. Biol. Chem.* **281,** 10263–10272 (2006).
38. Kanner, B. I. & Borre, L. The dual-function glutamate transporters: structure and molecular characterization of the substrate binding sites. *Biochim. Biophys. Acta* **1555,** 92–95 (2002).
39. Zarbiv, R., Grunewald, M., Kavanaugh, M. P. & Kanner, B. I. Cysteine scanning of the surroundings of an alkali-ion binding site of the glutamate transporter GLT-1 reveals a conformationally sensitive residue. *J. Biol. Chem.* **273,** 14231–14237 (1998).
40. Zhang, Y. & Kanner, B. I. Two serine residues of the glutamate transporter GLT-1 are crucial for coupling the fluxes of sodium and the neurotransmitter. *Proc. Natl Acad. Sci. USA* **96,** 1710–1715 (1999).
41. Brocke, L., Bendahan, A., Grunewald, M. & Kanner, B. I. Proximity of two oppositely oriented reentrant loops in the glutamate transporter GLT-1 indentified by paired cysteine mutagenesis. *J. Biol. Chem.* **277,** 3985–3992 (2002).
42. Gaillard, I., Slotboom, D. J., Knol, J., Lolkema, J. S. & Konings, W. N. Purification and reconstitution of the glutamate carrier GltT of the thermophilic bacterium *Bacillus stearothermophilus*. *Biochemistry* **35,** 6150–6156 (1996).
43. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
44. CCP4 Project. N. The *CCP4* suite: programs for protein crystallography. *Acta Crystallogr. D* **50,** 760–763 (1994).
45. Kleywegt, G. J. Use of non-crystallographic symmetry in protein structure refinement. *Acta Crystallogr. D* **52,** 842–857 (1996).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Mindell for support, and B. Hille and R. MacKinnon for constructive criticism. X-ray diffraction data were measured at beamlines X4A and X29 at the National Synchrotron Light Source and 8.2.2 at the Advanced Light Source. This work was supported by a National Research Service Award postdoctoral fellowship (D.Y.) and by the National Institutes of Health (E.G). E.G. is an Investigator with the Howard Hughes Medical Institute.

Author Information The coordinates for the lithium-bound native (NAT), TBOA-bound (TB) and sodium-bound (NA) states are deposited in the Protein Data Bank under accession codes 2NWL, 2NWW and 2NWX, respectively. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to E.G. (gouauxe@ohsu.edu).

ARTICLES

Basis for a ubiquitin-like protein thioester switch toggling E1–E2 affinity

Danny T. Huang[1,2], Harold W. Hunt[2], Min Zhuang[2,3], Melanie D. Ohi[4], James M. Holton[5] & Brenda A. Schulman[1,2,3]

Ubiquitin-like proteins (UBLs) are conjugated by dynamic E1–E2–E3 enzyme cascades. E1 enzymes activate UBLs by catalysing UBL carboxy-terminal adenylation, forming a covalent E1~UBL thioester intermediate, and generating a thioester-linked E2~UBL product, which must be released for subsequent reactions. Here we report the structural analysis of a trapped UBL activation complex for the human NEDD8 pathway, containing NEDD8's heterodimeric E1 (APPBP1–UBA3), two NEDD8s (one thioester-linked to E1, one noncovalently associated for adenylation), a catalytically inactive E2 (Ubc12), and MgATP. The results suggest that a thioester switch toggles E1–E2 affinities. Two E2 binding sites depend on NEDD8 being thioester-linked to E1. One is unmasked by a striking E1 conformational change. The other comes directly from the thioester-bound NEDD8. After NEDD8 transfer to E2, reversion to an alternate E1 conformation would facilitate release of the E2~NEDD8 thioester product. Thus, transferring the UBL's thioester linkage between successive conjugation enzymes can induce conformational changes and alter interaction networks to drive consecutive steps in UBL cascades.

Post-translational modification with ubiquitin-like proteins, such as ubiquitin, NEDD8 and SUMO, is an essential eukaryotic regulatory mechanism[1]. Ubiquitin, NEDD8 and other UBLs are covalently conjugated via their C termini to targets by related, but distinct cascades that involve the sequential actions of E1, E2 and E3 enzymes[2–12] (Fig. 1a). For clarity, we designate covalent complexes with a tilde (~), and noncovalent complexes with a hyphen (–). E1 enzymes activate UBLs through multiple steps. First, E1 binds ATP, Mg^{2+} and the UBL, and catalyses adenylation of the UBL's C terminus. Second, E1's catalytic cysteine attacks the UBL~adenylate, producing a covalent thioester-linkage between the E1's catalytic cysteine and the UBL's C terminus. The thioester-linked UBL will be transferred directly to an E2. However, before this transfer, the E1 binds another UBL molecule at the adenylation active site. Thus, during the activation cycle, the E1 binds two UBL molecules, each at a distinct site: UBL(T) is linked to the E1's catalytic cysteine via a thioester, and UBL(A) is bound noncovalently at the adenylation active site. Next, this doubly UBL-loaded E1 associates with an E2. Finally, a transthiolation reaction ensues whereby UBL(T) is transferred from the E1's catalytic cysteine to the E2's catalytic cysteine, the E2~UBL thioester product is released from E1, and the activation cycle continues for the noncovalently associated UBL(A) molecule. Consequently, the E1 cycles back and forth between the doubly UBL-loaded and singly UBL(A)-loaded forms as it binds each free E2 substrate and releases each E2~UBL product. Following the activation process, the E2~UBL complex typically associates with an E3, which facilitates UBL transfer to the target.

It is important to understand how the structural properties of different enzyme forms drive successive steps in conjugation. Recent structures of the NEDD8 and SUMO E1s, alone and in noncovalent singly loaded complexes with NEDD8(A) or SUMO(A), respectively, revealed similar overall domain orientations[13–15]. However, several previous studies suggest distinct structural properties for E1 and E2 forms involved in latter steps of UBL activation. First, E1s and E2s display different relative affinities for each other in their free and covalent thioester-linked enzyme~UBL states: free E1s display low affinity for E2s, doubly UBL-loaded E1s bind their free E2 substrates with high affinity, and E2~UBL thioester products are released from E1s[4,6,7]. Second, these differential affinities between distinct enzyme forms are required for progression of E1–E2–E3 cascades, because there is structural overlap between the E1 and E3 binding sites on E2s, and E2s cannot bind their E1 and E3 partners simultaneously[16–19]. Finally, upon docking the structure of a complex between E1 and E2 domains onto full-length structures of apo or singly UBL(A)-loaded E1s, an E2 would bind the opposite side and face away from the E1's catalytic cysteine[13–15,17,20]. Thus, significant conformational changes would be required to enable the E1 and E2 catalytic cysteines to face each other. This raises the questions of what roles different E1 conformations might play during the activation cycle, and what would drive E1 conformational changes. To understand the molecular switches influencing E1, E2 and UBL interactions, we determined the structure of a trapped activation complex for the NEDD8 pathway.

Trapped UBL activation complex structure

Wild-type activation complexes containing an E1, UBL(T), UBL(A) and an E2 are not stable, because UBL(T) is readily passed from the E1's catalytic cysteine to that of E2. Although an E2's catalytic cysteine is also essential for long-range allosteric changes in E1 structure[21], it was necessary to use a catalytically inactive E2 harbouring a cysteine-to-alanine mutation to trap an activation complex that would provide insights into intermolecular interactions among all the components. Using this approach, we determined the crystal structure of a trapped activation complex containing: APPBP1–MBP-UBA3 (NEDD8's heterodimeric E1, with UBA3 fused to the C terminus of the MBP crystallization tag), two NEDD8s (NEDD8(T) thioester-linked to UBA3's catalytic Cys216 and NEDD8(A) noncovalently associated at the adenylation active site), MgATP, and a catalytic cysteine-to-alanine (C111A) mutant of Ubc12 (NEDD8's E2). All MBP contacts are to regions of APPBP1, UBA3 and

[1]Howard Hughes Medical Institute, [2]Departments of Structural Biology and Genetics/Tumor Cell Biology, St Jude Children's Research Hospital, Memphis, Tennessee 38105, USA. [3]Interdisciplinary Program, University of Tennessee Health Sciences Center, Memphis, Tennessee 38163, USA. [4]Department of Cell Biology, Harvard Medical School, Boston, Massachusetts 02115, USA. [5]Physical Biosciences Division, Lawrence Berkeley National Laboratory University of California, Berkeley, California 94720, USA.

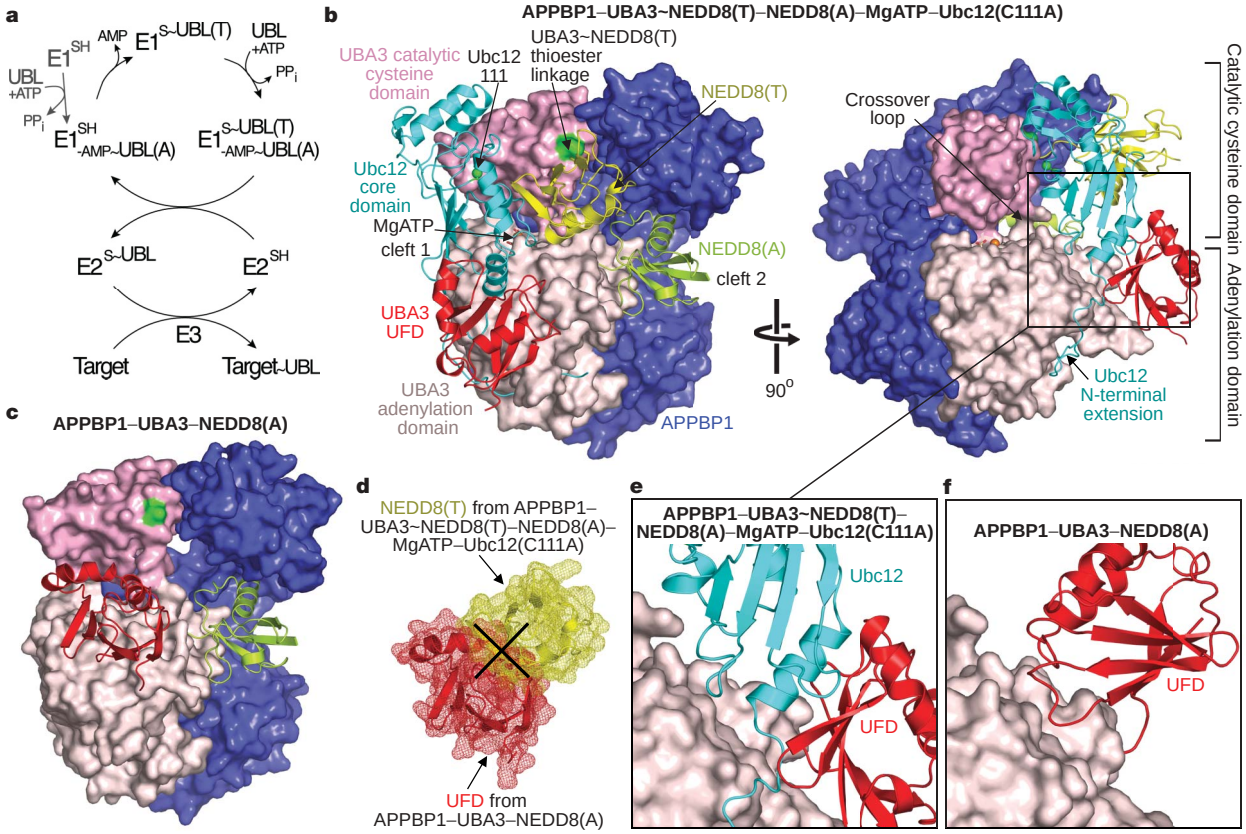


Figure 1 | Structure of APPBP1–UBA3~NEDD8(T)–NEDD8(A)–MgATP–Ubc12(C111A), a trapped UBL activation complex. a, UBL conjugation[2–12]. SH, free thiol of a catalytic cysteine; PP_i, inorganic pyrophosphate. **b**, The APPBP1–UBA3~NEDD8(T)–NEDD8(A)–MgATP–Ubc12(C111A) structure. Surfaces, APPBP1–UBA3's adenylation and cysteine domains; secondary structures, UBA3's UFD, NEDD8s and Ubc12. Blue, APPBP1; pink/red, UBA3; yellow, NEDD8(T); lime, NEDD8(A); cyan, Ubc12; green, UBA3's catalytic cysteine 216 and Ubc12's residue 111 (alanine here, but cysteine in wild-type Ubc12). **c**, APPBP1–UBA3–NEDD8(A)[14], coloured/oriented as is the left-side APPBP1–UBA3~NEDD8(T)–NEDD8(A)–MgATP–Ubc12(C111A) of **b**. **d**, APPBP1–UBA3~NEDD8(T)–NEDD8(A)–MgATP–Ubc12(C111A)'s NEDD8(T) overlaid with APPBP1–UBA3–NEDD8(A)'s UFD[14], coloured/oriented as in **b** and **c**. The cross shows structural overlap. **e**, Close-up of cryptic Ubc12-binding surface from APPBP1–UBA3~NEDD8(T)–NEDD8(A)–MgATP–Ubc12(C111A), oriented as is the right-side view in **b**. **f**, Close-up of region corresponding to **e** from APPBP1–UBA3–NEDD8(A)[14].

NEDD8(A) in conformations identical to previous structures, so MBP is not discussed further here. This trapped activation complex is referred to hereafter as APPBP1–UBA3~NEDD8(T)–NEDD8(A)–MgATP–Ubc12(C111A) (see Table 1, Supplementary Table 1 and Supplementary Fig. 1).

Previous structural studies revealed that APPBP1–UBA3 and other E1s display common modular architectures, with individual domains specifying each activity: an adenylation domain, a catalytic-cysteine-containing domain, and a domain structurally resembling ubiquitin (the ubiquitin-fold domain, UFD) that binds E2[13,15,17,22]. UBLs, such as ubiquitin and NEDD8, have two regions: an amino-terminal globular domain and a flexible C-terminal tail[23,24]. NEDD8's

Table 1 | X-ray refinement statistics

Resolution (Å)	50–2.8
R_{work}/R_{free}	0.241/0.274
Number of atoms	
Protein	13,012
Ligand/ion	33
Water	45
B-factors	
Protein	83.8
Ligand/ion	64.4
Water	59.4
Root-mean-square deviations	
Bond lengths (Å)	0.008
Bond angles (°)	1.44

$R_{work} = \sum|F_o - F_c|/\sum F_o$. R_{free} is the cross-validation of the *R*-factor for the test set of reflections (5% of total) omitted in model refinement.

E2, Ubc12, also has two regions: a unique N-terminal sequence, and a catalytic core domain conserved among all E2s[17,20,25]. These features are all visible in APPBP1–UBA3~NEDD8(T)–NEDD8(A)–MgATP–Ubc12(C111A), which adopts a compact overall structure. In this complex, the three APPBP1–UBA3 domains pack to generate a large central groove, which cradles the MgATP, both molecules of NEDD8, and Ubc12 substrates together (Fig. 1b). A crossover loop connecting the adenylation and catalytic-cysteine domains divides the groove into two clefts that are continuous both below and above the loop. As in previous structures, when viewed facing the E1 catalytic cysteine located centrally above the adenylation domain, NEDD8(A)'s globular domain binds in the right cleft (cleft 2) with its C-terminal tail extending under the crossover loop to approach ATP's α-phosphate in the left cleft (cleft 1)[13,14]. Ubc12's unique peptide-like extension docks in a groove unique to UBA3's adenylation domain, and Ubc12's core domain binds UBA3's UFD[17,20].

NEDD8(T) is in the centre of the complex, with its C terminus tethered within a channel focused on the thioester bond (Supplementary Fig. 2). A network of charged and polar side-chains contacts UBA3's catalytic Cys and NEDD8(T)'s C terminus. Mutational analysis shows that these residues contribute to APPBP1–UBA3~NEDD8(T) and Ubc12~NEDD8 complex formation. Conservation of this electrostatic network suggests that common mechanisms underlie E1-catalysed formation of E1~UBL and E2~UBL thioester complexes.

Relative to apo and singly UBL(A)-loaded E1 structures[13–15,20], the APPBP1–UBA3~NEDD8(T)–NEDD8(A)–MgATP–Ubc12(C111A)

structure reveals a striking ~120° rotation of the E2-binding UFD (Fig. 1, Supplementary Fig. 3). The UFD rotation results in remodelling of the APPBP1–UBA3 central groove to accommodate NEDD8(T)'s globular domain above the crossover loop in the middle of the groove, and Ubc12's core domain in cleft 1. In the alternative conformation found in previous apo and singly UBL(A)-loaded E1 structures, the central portion of the E1 groove is partially occupied by the E1's UFD[13–15,20] (Fig. 1c). With APPBP1–UBA3 doubly loaded with two NEDD8 molecules, the thioester-bound NEDD8(T) would clash with UBA3's UFD in the apo or singly UBL(A)-loaded orientation (Fig. 1d), suggesting that the UFD conformational change would accompany double NEDD8 loading of APPBP1-UBA3.

The large-scale UFD rotation reorients Ubc12 as compared to previous models such that (1) Ubc12's catalytic cysteine faces the direction of UBA3's catalytic cysteine; (2) Ubc12 is adjacent to the thioester-bound NEDD8(T); and (3) Ubc12 can bind two new E2-binding surfaces not present in apo and singly UBL(A)-loaded E1 forms. Interactions between doubly NEDD8-loaded APPBP1–UBA3 and Ubc12 bury 5,600 Å^2.

In the present structure, which represents a trapped rather than a functioning activation complex, there is a ~20 Å gap between Ubc12's residue 111 (here an alanine, but in wild-type Ubc12 the catalytic cysteine) and UBA3's catalytic cysteine 216. Although this gap is still greater than the distance required for transfer of NEDD8(T) to Ubc12, this finding is consistent with previous kinetic studies indicating that long-range conformational changes in E1 structure are induced by an E2's catalytic cysteine[21], which is absent from our structure. Accordingly, a corresponding ~20 Å gap in APPBP1–UBA3's central groove between NEDD8(T) and NEDD8(A) could accommodate further conformational changes to support the transthiolation reaction.

A cryptic Ubc12-binding site unmasked

One Ubc12-binding surface is a V-shaped groove. One face of the V comes from UBA3's UFD, and the other comes from the adenylation domain portion of UBA3, immediately adjacent to the ATP binding site (Figs 1b, e and 2a). In apo and singly UBL(A)-loaded E1s, this surface is concealed by the UFD in its alternative position[13–15,20] (Fig. 1f). However, in the structure containing two NEDD8s, the UFD has turned around and peeled away from the adjacent adenylation domain to create a single, continuous Ubc12-binding surface involving both domains. Mutational analysis underscores the importance of this cryptic site for E2 recruitment (Supplementary Fig. 4). Notably, the location of this Ubc12 binding site would allow transmission of subtle structural differences resulting from occupation of the nucleotide-binding site to the Ubc12 binding site.

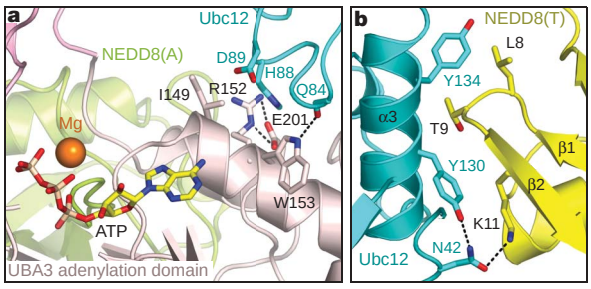


Figure 2 | Two Ubc12 binding sites depend on NEDD8(T) being thioester-linked to APPBP1–UBA3. a, Close-up view of the cryptic Ubc12-binding surface near the nucleotide-binding site, with the adenylation domain portion of UBA3 in pink, NEDD8(A) in lime, and Ubc12 in cyan. **b**, Close-up view of direct interactions between Ubc12 (cyan) and NEDD8(T) (yellow).

NEDD8(T) is also a binding site for Ubc12

Ubc12 also binds directly to NEDD8(T). A hydrophobic groove formed by Ubc12's α3 helix wraps around NEDD8(T)'s β1β2-loop, and additional electrostatic interactions further stabilize the complex (Fig. 2b). Other UBL pathways probably involve parallel interactions recruiting E2s to their corresponding doubly UBL-loaded E1 complexes, because structural studies of thioester analogue complexes between ubiquitin and SUMO with their E2s also reveal interactions between UBL surfaces corresponding to NEDD8's β1β2-loop and E2 surfaces corresponding to Ubc12's α3 helix[19,26–28].

Alanine-scanning over Ubc12's core domain revealed that mutations at each interaction surface significantly decrease formation of the Ubc12~NEDD8 thioester product (Supplementary Fig. 5). Further enzyme kinetic analyses revealed increased Michaelis constant (K_m) values for mutants with substitutions for residues involved in binding (1) UBA3's UFD, (2) UBA3's adenylation domain adjacent to the ATP-binding site, and (3) the thioester-bound NEDD8(T) (Supplementary Table 2). Thus, the structurally observed E2-binding surfaces are critical for Ubc12 recruitment for the transthiolation reaction.

Model for E1–E2 transthiolation

Our structure suggests a model for transthiolation, in which the E1 and E2 cysteine-to-cysteine gap would be closed. Continuing along the same trajectory as between the apo or singly NEDD8(A)-loaded E1 and the trapped activation complex presented here, an additional ~10° rotation of UBA3's UFD would bring the E1 and E2 cysteines into close proximity (Fig. 3a). Because E2 core domains have oblong, tower-like structures, a small rotation at the base translates into a

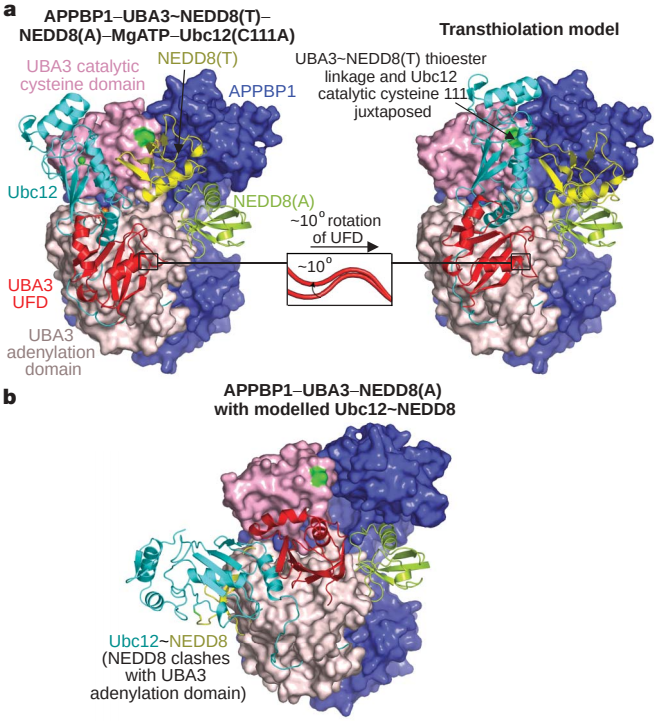


Figure 3 | Model of a transthiolation complex. a, An additional ~10° UFD rotation allows juxtaposition of Ubc12's catalytic Cys111 with the UBA3~NEDD8(T) thioester, in a putative conformation for the transthiolation reaction. The APPBP1–UBA3~NEDD8(T)–NEDD8(A)–MgATP–Ubc12(C111A) structure is in the left panel, the model is in the right panel, and proteins are coloured as in Fig. 1b. **b**, Superposition of the UFD–Ubc12~NEDD8 thioester model onto the APPBP1–UBA3–NEDD8(A) structure[14], generated by least-squares alignment of Cα atoms in UBA3's UFD. Much of NEDD8 in the Ubc12~NEDD8 thioester model is not visible owing to clashing with the APPBP1–UBA3 surface.

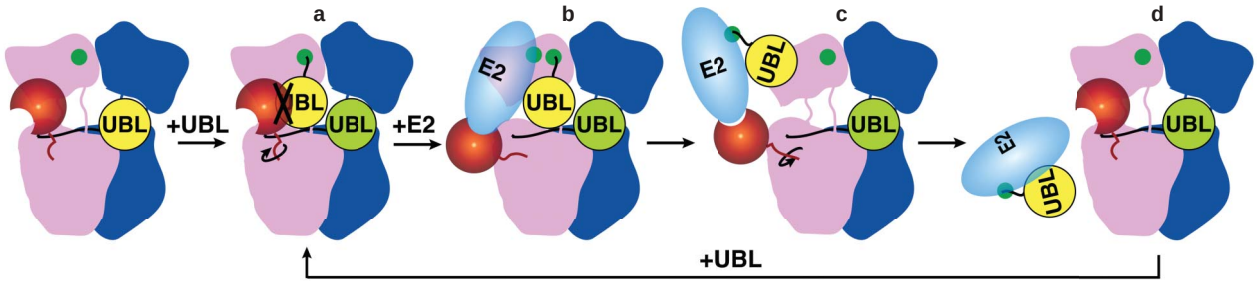


Figure 4 | A thioester switch toggling E1–E2 interactions. E1 is blue/pink, corresponding to APPBP1/UBA3, with UBA3's UFD red. E2 is cyan. The first and second UBLs binding E1 are yellow and lime, respectively. Catalytic cysteines are green. **a**, UBL(T), thioester-bound to E1, clashes with E1's UFD in initial conformation. **b**, E1's UFD rotation unmasks cryptic E2-binding sites, allowing doubly UBL-loaded E1 to bind free E2. **c**, UBL transfer to E2's cysteine eliminates the UBL's covalent tether to E1. This removes E2-binding sites, and allows reversion to the alternative E1 conformation. **d**, Steric clashing between E1 and E2~UBL facilitates product release. Another activation cycle ensues.

large translation for the catalytic cysteine. Several features of our analyses support this model. First, this rotation would also require a corresponding movement of NEDD8(T), and there is an equivalent gap in APPBP1–UBA3's central groove to accommodate NEDD8(T) in a rotated position without steric clashes. Second, other than interactions with Ubc12, NEDD8(T) is tethered to APPBP1–UBA3 through its flexible C terminus. The structural flexibility of UBL C-terminal tails[23,24] would allow NEDD8(T) to rotate along with Ubc12. Finally, our model predicts one additional Ubc12 surface involved in transthiolation: the channel leading to Ubc12's catalytic cysteine 111 would secure NEDD8(T)'s C-terminal tail for the reaction. In support of the model, our alanine-scanning revealed that residues lining the channel to cysteine 111 contribute the only side-chains involved in forming the Ubc12~NEDD8 thioester complex that are outside the structurally observed interfaces (Supplementary Fig. 5). It seems likely that Ubc12's catalytic cysteine, absent from our trapped complex, stabilizes the rotated, active conformation[21]. It is also possible that adenylation of the noncovalently bound NEDD8(A) may trigger conformational changes in UBA3's adenylation domain, contributing to rotation at the base of Ubc12.

After the transthiolation reaction, the Ubc12~NEDD8 thioester product is released. Elimination of the covalent tether between NEDD8(T) and UBA3 would allow rotation of UBA3's UFD back to the conformation observed in the apo and singly NEDD8(A)-loaded APPBP1–UBA3 structures. To gain insight into this process, we docked the Ubc12~NEDD8 thioester complex from our transthiolation model onto the previous APPBP1–UBA3–NEDD8(A) structure[14]. With the UFD in the alternate orientation, substantial clashing between Ubc12~NEDD8 and APPBP1–UBA3 could facilitate product release (Fig. 3b). Similarity in the overall domain orientations in the apo and singly UBL(A)-loaded structures of SUMO's E1[15] (Supplementary Fig. 3) indicates that UFD rotations have similar roles in the activation cycles of other UBLs.

A thioester switch modulates affinities

Our results suggest that a UBL-dependent thioester switch toggles E1–E2 interactions to drive the activation cycle (Fig. 4). First, E1 is switched into a form that binds E2 (Fig. 4a, b). In the doubly UBL-loaded E1, with UBL(T) thioester-linked to E1, the E1's conformational change reorients the E2-binding UFD and unmasks a cryptic E2-binding site. Another E2-binding site comes directly from UBL(T) thioester-linked to E1. Transfer of UBL(T)'s covalent linkage from E1 to the E2's cysteine generates the E2~UBL thioester product. In the absence of UBL(T)'s covalent tether to E1, the switch flips back (Fig. 4c). Lack of UBL(T)-dependent E2-binding sites, combined with steric clashing between the E2~UBL thioester product and the singly UBL(A)-loaded E1 is likely to facilitate product release (Fig. 4d). Thus, E1–E2 interactions depend on whether E1 or E2 is thioester-bound to the UBL. Formation of the E1~UBL thioester favours interaction with free E2, whereas formation of the E2~UBL thioester leads to release.

Implications for UBL cascades

A fundamental characteristic of UBL pathways is the sequential formation of transient complexes during E1–E2–E3 conjugation cascades, and also during molecular 'handoffs' of monoubiquitinated cargo between different ubiquitin-recognition complexes. The same surfaces of enzymes, and of UBLs, are recognized at consecutive steps during many stages of these pathways[18,29–34]. For UBL pathways to proceed, distinct transient interactions and/or catalytic activities are probably specified by different states of the E1–E2–E3 enzymes, and of UBL recognition machineries. Our results show how covalent tethering of a UBL to an enzyme's active site can allow switching between states, with altered protein–protein interaction abilities, that are required for progression of the transfer cascade. Notably, NEDD8 and UBA3's UFD, which occupy overlapping positions in the different E1 states, are structurally similar. Many other proteins in UBL pathways also contain domains structurally resembling UBLs that could play roles in switching between states. It is likely that the UBL handoffs occurring at other stages of conjugation and recognition pathways are likewise driven forward by each reaction inducing conformational changes and altering interaction networks to trigger the next step in the cascade.

METHODS

Protein preparation, crystallization, structure determination and biochemical assays are described in detail in the Supplementary Information.

APPBP1–UBA3~NEDD8(T)–NEDD8(A)–MgATP–Ubc12(C111A) was generated by mixing purified APPBP1–MBP-Δ11UBA3, Δ5-2Ubc12(Cys111Ala), and NEDD8 at a 1:2:4 molar ratio in 25 mM HEPES, 150 mM NaCl, 5 mM $MgCl_2$, 5 mM ATP (pH 7.0) at 4 °C for 18 h, and purified by gel filtration in 25 mM HEPES, 150 mM NaCl, 0.5 mM $MgCl_2$, 0.5 mM ATP (pH 7.0). Crystals grew immediately upon setting drops at 18 °C by the hanging-drop vapour diffusion method, in 17% PEG 3350, 0.2 M di-sodium tartrate, 0.1 M HEPES (pH 7.0). The crystals belong to space group $P3_221$ with $a = b = 156.5$ Å, $c = 190.5$ Å, and contain one complex per asymmetric unit.

Diffraction data were collected at the APS SERCAT and the ALS 8.2.2 and 8.3.1 beamlines. The structure was determined by molecular replacement, with four search models: (1) the APPBP1–UBA3 adenylation domain (APPBP1's residues 9–167 and 395–534 and UBA3's residues 12–208 and 293–347); (2) the APPBP1–UBA3 catalytic cysteine domain (APPBP1's residues 181–374 and UBA3's residues 213–287); (3) NEDD8 residues 1–73; and (4) the complex between UBA3's UFD and Ubc12's core domain[13,14,17,20]. APPBP1's residues 1–4, MBP's residues 1, 299–302, and 312–319, and Ubc12's residues 1–2 are presumably disordered and are not present in the final model, which was refined to 2.8 Å resolution and $R = 24.1$, $R_{free} = 27.4$.

Received 6 October; accepted 27 November 2006.
Published online 14 January 2007.

1. Kerscher, O., Felberbaum, R. & Hochstrasser, M. Modification of proteins by ubiquitin and ubiquitin-like proteins. *Annu. Rev. Cell Dev. Biol.* **22**, 159–180 (2006).

2. Haas, A. L., Warms, J. V., Hershko, A. & Rose, I. A. Ubiquitin-activating enzyme. Mechanism and role in protein-ubiquitin conjugation. *J. Biol. Chem.* **257,** 2543–2548 (1982).
3. Haas, A. L. & Rose, I. A. The mechanism of ubiquitin activating enzyme. A kinetic and equilibrium analysis. *J. Biol. Chem.* **257,** 10329–10337 (1982).
4. Haas, A. L., Bright, P. M. & Jackson, V. E. Functional diversity among putative E2 isozymes in the mechanism of ubiquitin-histone ligation. *J. Biol. Chem.* **263,** 13268–13275 (1988).
5. Ciechanover, A., Elias, S., Heller, H. & Hershko, A. "Covalent affinity" purification of ubiquitin-activating enzyme. *J. Biol. Chem.* **257,** 2537–2542 (1982).
6. Hershko, A., Heller, H., Elias, S. & Ciechanover, A. Components of ubiquitin-protein ligase system. Resolution, affinity purification, and role in protein breakdown. *J. Biol. Chem.* **258,** 8206–8214 (1983).
7. Pickart, C. M. & Rose, I. A. Functional heterogeneity of ubiquitin carrier proteins. *J. Biol. Chem.* **260,** 1573–1581 (1985).
8. Pickart, C. M., Kasperek, E. M., Beal, R. & Kim, A. Substrate properties of site-specific mutant ubiquitin protein (G76A) reveal unexpected mechanistic features of ubiquitin-activating enzyme (E1). *J. Biol. Chem.* **269,** 7115–7123 (1994).
9. Hershko, A. & Ciechanover, A. The ubiquitin system. *Annu. Rev. Biochem.* **67,** 425–479 (1998).
10. Hochstrasser, M. Biochemistry. All in the ubiquitin family. *Science* **289,** 563–564 (2000).
11. Bohnsack, R. N. & Haas, A. L. Conservation in the mechanism of Nedd8 activation by the human AppBp1-Uba3 heterodimer. *J. Biol. Chem.* **278,** 26823–26830 (2003).
12. Pickart, C. M. & Eddins, M. J. Ubiquitin: structures, functions, mechanisms. *Biochim. Biophys. Acta* **1695,** 55–72 (2004).
13. Walden, H., Podgorski, M. S. & Schulman, B. A. Insights into the ubiquitin transfer cascade from the structure of the activating enzyme for NEDD8. *Nature* **422,** 330–334 (2003).
14. Walden, H. *et al.* The structure of the APPBP1–UBA3-NEDD8-ATP complex reveals the basis for selective ubiquitin-like protein activation by an E1. *Mol. Cell* **12,** 1427–1437 (2003).
15. Lois, L. M. & Lima, C. D. Structures of the SUMO E1 provide mechanistic insights into SUMO activation and E2 recruitment to E1. *EMBO J.* **24,** 439–451 (2005).
16. Bencsath, K. P., Podgorski, M. S., Pagala, V. R., Slaughter, C. A. & Schulman, B. A. Identification of a multifunctional binding site on Ubc9p required for Smt3p conjugation. *J. Biol. Chem.* **277,** 47938–47945 (2002).
17. Huang, D. T. *et al.* Structural basis for recruitment of Ubc12 by an E2 binding domain in NEDD8's E1. *Mol. Cell* **17,** 341–350 (2005).
18. Eletr, Z. M., Huang, D. T., Duda, D. M., Schulman, B. A. & Kuhlman, B. E2 conjugating enzymes must disengage from their E1 enzymes before E3-dependent ubiquitin and ubiquitin-like transfer. *Nature Struct. Mol. Biol.* **12,** 933–934 (2005).
19. Reverter, D. & Lima, C. D. Insights into E3 ligase activity revealed by a SUMO–RanGAP1–Ubc9–Nup358 complex. *Nature* **435,** 687–692 (2005).
20. Huang, D. T. *et al.* A unique E1–E2 interaction required for optimal conjugation of the ubiquitin-like protein NEDD8. *Nature Struct. Mol. Biol.* **11,** 927–935 (2004).
21. Tokgoz, Z., Bohnsack, R. N. & Haas, A. L. Pleiotropic effects of ATP.Mg^{2+} binding in the catalytic cycle of ubiquitin-activating enzyme. *J. Biol. Chem.* **281,** 14729–14737 (2006).
22. Szczepanowski, R. H., Filipek, R. & Bochtler, M. Crystal structure of a fragment of mouse ubiquitin-activating enzyme. *J. Biol. Chem.* **280,** 22006–22011 (2005).
23. Vijay-Kumar, S., Bugg, C. E. & Cook, W. J. Structure of ubiquitin refined at 1.8 Å resolution. *J. Mol. Biol.* **194,** 531–544 (1987).
24. Whitby, F. G., Xia, G., Pickart, C. M. & Hill, C. P. Crystal structure of the human ubiquitin-like protein NEDD8 and interactions with ubiquitin pathway enzymes. *J. Biol. Chem.* **273,** 34983–34991 (1998).
25. Osaka, F. *et al.* A new NEDD8-ligating system for cullin-4A. *Genes Dev.* **12,** 2263–2268 (1998).
26. Miura, T., Klaus, W., Gsell, B., Miyamoto, C. & Senn, H. Characterization of the binding interface between ubiquitin and class I human ubiquitin-conjugating enzyme 2b by multidimensional heteronuclear NMR spectroscopy in solution. *J. Mol. Biol.* **290,** 213–228 (1999).
27. Hamilton, K. S. *et al.* Structure of a conjugating enzyme-ubiquitin thiolester intermediate reveals a novel role for the ubiquitin tail. *Structure* **9,** 897–904 (2001).
28. Brzovic, P. S., Lissounov, A., Christensen, D. E., Hoyt, D. W. & Klevit, R. E. A UbcH5/ubiquitin noncovalent complex is required for processive BRCA1-directed ubiquitination. *Mol. Cell* **21,** 873–880 (2006).
29. Hicke, L., Schubert, H. L. & Hill, C. P. Ubiquitin-binding domains. *Nature Rev. Mol. Cell Biol.* **6,** 610–621 (2005).
30. Harper, J. W. & Schulman, B. A. Structural complexity in ubiquitin recognition. *Cell* **124,** 1133–1136 (2006).
31. Hurley, J. H. & Emr, S. D. The ESCRT complexes: structure and mechanism of a membrane-trafficking network. *Annu. Rev. Biophys. Biomol. Struct.* **35,** 277–298 (2006).
32. Hicke, L. A new ticket for entry into budding vesicles—ubiquitin. *Cell* **106,** 527–530 (2001).
33. Morita, E. & Sundquist, W. I. Retrovirus budding. *Annu. Rev. Cell Dev. Biol.* **20,** 395–425 (2004).
34. Hochstrasser, M. Lingering mysteries of ubiquitin-chain assembly. *Cell* **124,** 27–34 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We dedicate this manuscript to the memory of C. Pickart for her contributions to our understanding of ubiquitin and UBL pathways, and for advice and encouragement throughout the course of these studies. We thank C. Ralston, B. Sankaran and A. Howard for assistance with data collection at the 8.2.2 beamline at ALS and at the SERCAT beamline at APS. We thank P. Murray, D. Minor, B. Dye and D. Scott for critical reading of the manuscript, members of the Schulman laboratory for discussions, and C. Ross for X-ray support. This work was supported in part by ALSAC and grants from the NIH (B.A.S.) and the Charles A. King Medical Foundation (M.D.O.). B.A.S. is an Investigator of the Howard Hughes Medical Institute.

Author Contributions D.T.H. and B.A.S. designed the experiments. D.T.H., H.W.H., M.Z., J.M.H. and B.A.S. performed the experiments. M.D.O. assessed the quality of crystallization samples and made conceptual contributions. D.T.H. and B.A.S. wrote the manuscript.

Author Information Coordinates and structure factors are deposited in the Protein Data Bank under accession code 2NVU. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to B.A.S. (brenda.schulman@stjude.org).

LETTERS

An unexpected cooling effect in Saturn's upper atmosphere

C. G. A. Smith[1], A. D. Aylward[1], G. H. Millward[1]†, S. Miller[1] & L. E. Moore[2]

The upper atmospheres of the four Solar System giant planets exhibit high temperatures[1,2] that cannot be explained by the absorption of sunlight[2,3]. In the case of Saturn the temperatures predicted by models of solar heating[2,4] are ~200 K, compared to temperatures of ~400 K observed independently in the polar regions[5] and at 30° latitude[6]. This unexplained 'energy crisis' represents a major gap in our understanding of these planets' atmospheres. An important candidate for the source of the missing energy is the magnetosphere[1,2,4,7–9], which injects energy mostly in the polar regions of the planet. This polar energy input is believed to be sufficient to explain the observed temperatures[9], provided that it is efficiently redistributed globally by winds[4,8], a process that is not well understood. Here we show, using a numerical model[4], that the net effect of the winds driven by the polar energy inputs is not to heat but to cool the low-latitude thermosphere. This surprising result allows us to rule out known polar energy inputs as the solution to the energy crisis at Saturn. There is either an unknown—and large—source of polar energy, or, more probably, some other process heats low latitudes directly.

Recent numerical modelling studies[4,8] have shown that under specific circumstances polar energy inputs may explain the high thermospheric temperatures at Saturn. For plausible heating distributions in the polar regions there is predicted to exist a system of equatorward winds that redistribute the energy globally, generating the observed temperatures at both low and high latitudes. But these studies consider only the effects of pure thermal energy inputs on the thermosphere, whereas the bulk of the polar energy input from the magnetosphere is thought to be a mixture of thermal energy (Joule heating) and kinetic energy (ion drag) in roughly equal proportions[7,9]. The kinetic energy input is necessarily accompanied by an input of angular momentum. The sense of this angular momentum input is westward, that is, in the opposite direction to that of the planetary rotation. The overall effect of ion drag on the dynamics is thus expected to be the generation of strong westward winds throughout the polar thermosphere. Here we address for the first time to our knowledge the effect of these westward winds on the structure of the thermosphere.

We use a numerical thermosphere model[4] based on a widely used model of the terrestrial thermosphere[10], converted to an atmosphere composed of H_2, H and He. It uses an eulerian grid on which temperatures and the three components of the neutral wind are calculated by time integration. The lower boundary is placed at an altitude of 800 km above the 1 bar level, and at a fixed pressure of 100 nbar. At this level we assume a fixed temperature of 143 K (ref. 11) and a horizontal wind velocity of zero. We parameterize the vertical transport of energy and momentum by small-scale motions using an eddy diffusion coefficient[12] of $K_\tau = 10^4\ m^2\ s^{-1}$. For this study, we simplify the model by assuming that the system is symmetric about the planet's axis of rotation and mirror-symmetric about the equatorial plane. These are good approximations because Saturn's magnetic field almost exhibits both of these symmetries[13]. Furthermore, the dynamics of the magnetosphere are strongly dominated by the planetary rotation and can be approximated as axially symmetric to first order[9]. We expect deviations from our assumed symmetries to be second-order effects. We note that the introduction of axial symmetry does not mean that we do not model zonal (east–west) winds: the effects of zonal winds are fully included in the model, but they are assumed to be identical at all longitudes. The introduction of these symmetry assumptions allows us to use very high grid resolutions in our model. We employ a latitudinal resolution of 0.2° and a vertical resolution of 0.2 pressure scale heights. Further details and discussion of the model are given in the Supplementary Methods.

This basic model is forced only by solar heating. Running this model to near steady state (requiring a run time of 400 planetary rotations[8]) predicts roughly uniform global temperatures of ~150–160 K (Supplementary Fig. 1), with the higher temperatures at the equator. To include the polar energy inputs, we require models of the ionospheric conductivity and plasma flows. Good empirical models of the ionospheric conductivity do not exist, owing to lack of data. For this reason we use a conductivity distribution calculated using a numerical model of the ionosphere[14]. We use ion and electron densities from this model to calculate a global distribution of conductivity and fix these values with respect to our thermosphere model. The ionospheric plasma flows are taken from an empirical model[9] based on a mixture of *in situ* spacecraft[15] and ground-based spectroscopic data[16]. This model predicts that the magnetosphere will exert a westward ion drag on the thermosphere poleward of ~65° latitude. Thus we expect the Joule heating and ion drag to be significant only in this region. Further details of both these models, and our formulation of Joule heating and ion drag, are given in the Supplementary Methods.

Including these inputs in the model generates the structures shown in Fig. 1. Comparing this run to the control run forced only by solar heating shows that Joule heating and ion drag cause net heating poleward of and net cooling equatorward of the dashed line. The maximum cooling effect is ~7 K at ~65° latitude. Although small, this effect is unexpected, counter-intuitive and rather surprising. It can be explained by inspecting the meridional circulation, which shows a poleward flow at low altitudes and an equatorward flow at high altitudes. The low-altitude flow is energetically the more important because the density of the atmosphere decreases with increasing altitude. Thus the dynamical coupling between low and high latitudes is dominated by a steady flow of gas—and therefore energy—away from low latitudes and into the polar regions. This acts both to enhance convective cooling of low latitudes, and to heat the polar region. The flows themselves arise because of small force imbalances in the thermosphere. In Fig. 2 we show a conceptual

[1]Department of Physics and Astronomy, University College London, WC1E 6BT, UK. [2]Center for Space Physics, Boston University, Boston, Massachusetts 02215, USA. †Present address: Laboratory for Atmospheric and Space Physics, University of Colorado, Boulder, Colorado 80303, USA.

interpretation of these flows in terms of the hydrostatic balance in the upper atmosphere; in the Supplementary Discussion we also present an analysis of the actual forces calculated by the model.

It is clear from Fig. 1 that, while our model does reproduce the observed temperature of ~400 K in the polar regions[5], we do not reproduce the observed value of ~400 K at 30° latitude[6], because we predict cooling everywhere equatorward of 55° latitude. The initial indication is thus that the polar energy inputs resolve the high-latitude energy crisis but do not resolve the low-latitude energy crisis. However, we must consider whether our results depend on our choice of model inputs. The input in which we have the least confidence is the ionospheric conductivity model: while our thermospheric boundary conditions and plasma flow model are well supported by the available data, the conductivity model is not. In particular, the model currently only includes ionization due to absorption of solar radiation, and so it most probably underestimates the conductivity in the polar regions, where particle precipitation is an important source of ionization. We have thus performed a sensitivity study in which we artificially scale up the conductivity globally by fixed factors. The effect of this alteration is simply to intensify the behaviour described above. If the scaling factor is 16, the polar hotspot reaches temperatures greater than 500 K and the maximum degree of cooling compared to the control run increases to ~27 K (see the Supplementary Discussion for further details). Thus we still approximately match the observed high-latitude temperature and still fail to match the observed low-latitude temperature.

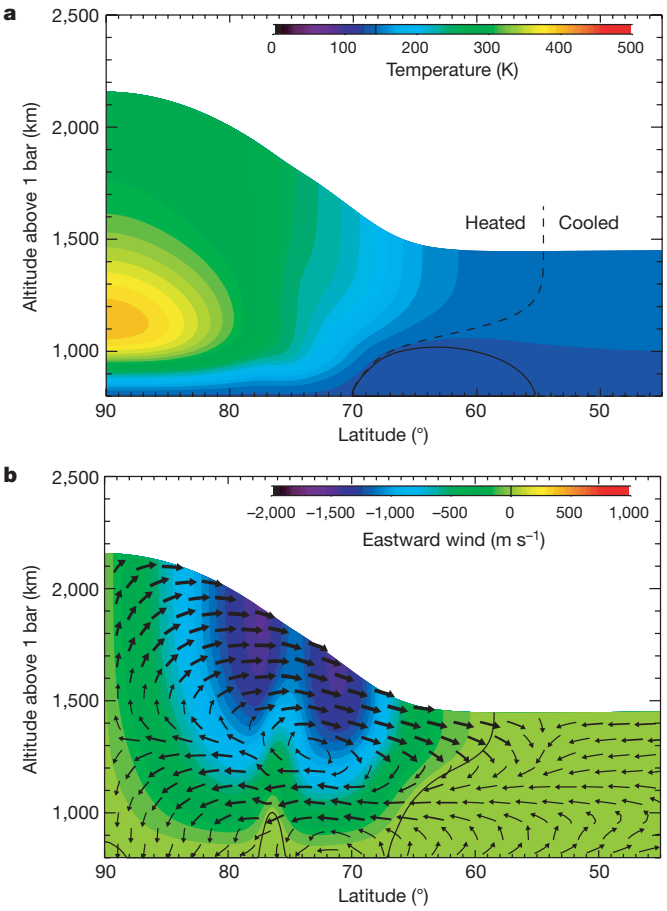


Figure 1 | Thermal and dynamical structure of the upper atmosphere predicted by our model. **a**, Temperature structure (colour contours). The dashed line separates regions to the left, which are heated by the polar energy inputs, from regions on the right, which are cooled. At latitudes smaller than those shown the temperature profile remains approximately constant to the equator, exhibiting exospheric temperatures of 150–160 K. It is thus clear that the polar energy inputs do not reproduce the observed temperature of ~400 K at 30° latitude[6]. However, the temperature of ~400 K at the pole is a good match to the infrared spectroscopic temperatures determined in this region[5]. The solid contour represents the fixed lower-boundary temperature of 143 K. The region enclosed by this contour at ~55–70° latitude is thus cooler than the lower-boundary temperature. **b**, Zonal winds (colour contours). The zero contour of zonal wind is shown by the solid line. Poleward of ~65° latitude the winds are almost entirely westward, as expected from the direction of ion drag. The double-lobed structure in the zonal winds is due to structures in the plasma flow model (see the Supplementary Information). Arrows show the combined vertical and meridional circulation. The thinnest arrows represent wind speeds of $<1\ m\ s^{-1}$, the thickness increasing linearly with the logarithm of wind speed until the thickest arrows represent wind speeds of $>100\ m\ s^{-1}$. The cooling effect is produced by two mechanisms. First, the poleward flow induced by ion drag enhances the convective cooling of the low-latitude regions. Second, the increase in the poleward wind speed between 55° and 70° latitude represents a divergence in the flow that must be balanced by upwelling to satisfy continuity. This upwelling gas expands, and cools adiabatically. It is this effect that produces temperatures cooler than the lower-boundary temperature in this region.

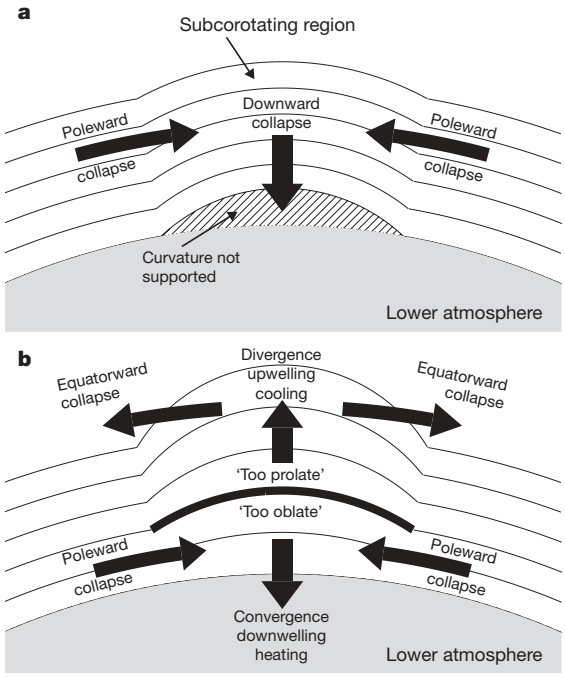


Figure 2 | Interpretation of polar dynamics in terms of hydrostatic balance. For the atmosphere to rest in hydrostatic balance, we require internal pressure gradients to perfectly balance the combined gravitational and centrifugal accelerations. For this reason, surfaces of constant pressure normally align themselves with surfaces of constant potential energy: this is the cause of Saturn's considerable oblateness. If the polar upper atmosphere is made to rotate more slowly by ion drag, but in the same gravity field, it must adopt a less-oblate profile if it is to rest in perfect hydrostatic balance. In **a**, we show the situation that this implies if the atmosphere is isothermal. Here the grey shaded region represents the atmosphere lying below the region that we model and the solid lines represent surfaces of constant pressure in the upper atmosphere. In the subcorotating region close to the pole, the less-oblate profile that we require is not supported by the lower atmosphere, and the gas thus 'collapses' towards the pole and downwards. This inward flow is strongly convergent, causing downwelling and compression that heats the gas. The steady-state situation is sketched in **b**. Here, the compressional heating at the pole has increased the scale height, allowing a less-oblate profile to be supported. However, owing to the efficiency of vertical thermal conduction, this increased scale height persists at all altitudes. Thus, there is a particular pressure surface for which the curvature is 'just right' to support the rotation velocity of the gas (thick line). At lower altitudes the curvature is 'too oblate' to support the rotation, and the gas continues to collapse inwards and provide compressional heating. At higher altitudes the curvature is 'too prolate' and the gas collapses away from the pole.

We must also ask whether we are omitting any important sources of polar energy. The only well-quantified source of polar energy other than those used in our model is energy deposition by the particle precipitation that forms the ultraviolet aurora[9,17]. This has already been shown to have a negligible effect on the thermal structure[8]. It has also been suggested that small-scale structures in the ionospheric plasma flow may greatly increase the total Joule heating[7,18] without a corresponding increase in ion drag. Such structures would presumably originate in the magnetosphere and solar wind. Data collected in these regions by the Cassini mission may allow quantification of this small-scale structure and its implications for the flow of energy within the magnetosphere–atmosphere system.

However, the good match between our results and the high-latitude temperature measurements[5] suggests that if we were to introduce additional polar energy inputs that were sufficient to outweigh the cooling effect and thus resolve the low-latitude energy crisis, we would probably overheat the polar regions. Thus our results strongly suggest that low latitudes are heated directly, perhaps by the breaking of buoyancy waves generated in the lower atmosphere[19–22]. There may also be some Joule heating or particle precipitation at low latitudes that has yet to be accounted for. The Cassini mission may contribute to an improved understanding of such processes by providing new measurements of the thermospheric temperature. Although a number of low-latitude temperature measurements are at present available, they are neither mutually consistent nor unambiguous[6,23–25]. A multi-latitude thermospheric data set collected by Cassini and analysed self-consistently would thus be an invaluable resource.

In summary, our conclusions indicate strongly that polar energy inputs are not the solution to the low-latitude energy crisis at Saturn, and that future research should thus focus on direct heating of low latitudes. We expect our results to apply in outline to the slightly more complicated situation at Jupiter, and preliminary results from a jovian version of our model support this prediction. However, we are not yet in a position to assess whether our cooling effect may be relevant to the energy crises at Uranus and Neptune, given the apparent complexity of their magnetospheres, but this study does indicate that magnetosphere–atmosphere coupling at these planets is likely to be complicated and may throw up further surprises.

Received 11 July; accepted 8 December 2006.

1. Atreya, S. K. *Atmosphere and Ionospheres of the Outer Planets and their Satellites* Ch. 2 (Springer, Heidelberg, 1986).
2. Yelle, R. V. & Miller, S. in *Jupiter: Planet, Satellites and Magnetosphere* (eds Bagenal, F., McKinnon, W. & Dowling, T.) 185–218 (Cambridge Univ. Press, Cambridge, UK, 2004).
3. Strobel, D. F. & Smith, G. R. On the temperature of the jovian thermosphere. *J. Atmos. Sci.* **30,** 718–725 (1973).
4. Mueller-Wodarg, I. C. F., Mendillo, M., Yelle, R. V. & Aylward, A. D. A global circulation model of Saturn's thermosphere. *Icarus* **180,** 147–160 (2006).
5. Melin, H. *Comparative Aeronomy of the Upper Atmospheres of the Giant Planets.* PhD thesis, Univ. London (2006).
6. Smith, G. R. *et al.* Saturn's upper atmosphere from the Voyager 2 EUV solar and stellar occultations. *J. Geophys. Res.* **88,** 8667–8679 (1983).
7. Smith, C. G. A., Miller, S. & Aylward, A. D. Magnetospheric energy inputs into the upper atmospheres of the giant planets. *Ann. Geophys.* **23,** 1943–1947 (2005).
8. Smith, C. G. A., Aylward, A. D., Miller, S. & Mueller-Wodarg, I. C. F. Polar heating in Saturn's thermosphere. *Ann. Geophys.* **23,** 2465–2477 (2005).
9. Cowley, S. W. H., Bunce, E. J. & O'Rourke, J. M. A simple quantitative model of plasma flows and currents in Saturn's polar ionosphere. *J. Geophys. Res.* **A18,** 5212–5230 (2004).
10. Fuller-Rowell, T. J. *et al.* in *STEP Handbook of Ionospheric Models* 217–238 (SCOSTEP, Logan, Utah, 1996).
11. Moses, J. I. *et al.* Photochemistry of Saturn's atmosphere. I. Hydrocarbon chemistry and comparisons with ISO observations. *Icarus* **143,** 244–298 (2000).
12. Atreya, S. K. Eddy mixing coefficient on Saturn. *Planet. Space Sci.* **30,** 849–854 (1982).
13. Davis, L. J. & Smith, E. J. A model of Saturn's magnetic field based on all available data. *J. Geophys. Res.* **95,** 15257–15261 (1990).
14. Moore, L. E., Mendillo, M., Mueller-Wodarg, I. C. F. & Murr, D. L. Modeling of global variations and ring shadowing in Saturn's ionosphere. *Icarus* **172,** 503–520 (2004).
15. Richardson, J. D. Thermal ions and Saturn—Plasma parameters and implications. *J. Geophys. Res.* **91,** 1381–1389 (1986).
16. Stallard, T. S., Miller, S., Trafton, L. M., Geballe, T. R. & Joseph, R. D. Ion winds in Saturn's southern auroral/polar region. *Icarus* **167,** 204–211 (2004).
17. Clarke, J. T. *et al.* Morphological differences between Saturn's ultraviolet aurorae and those of Earth and Jupiter. *Nature* **433,** 717–719 (2005).
18. Codrescu, M. V., Fuller-Rowell, T. J. & Foster, J. C. On the importance of E-field variability for Joule heating in the high-latitude thermosphere. *Geophys. Res. Lett.* **22,** 2393–2396 (1995).
19. Young, L. A., Yelle, R. V., Young, R., Seiff, A. & Kirk, D. B. Gravity waves in Jupiter's thermosphere. *Science* **276,** 108–111 (1997).
20. Matcheva, K. I. & Strobel, D. F. Heating of Jupiter's thermosphere by dissipation of gravity waves due to molecular viscosity and heat conduction. *Icarus* **140,** 328–340 (1999).
21. Hickey, M. P., Walterscheid, R. L. & Schubert, G. Gravity wave heating and cooling in Jupiter's thermosphere. *Icarus* **148,** 266–281 (2000).
22. Hickey, M. P., Schubert, G. & Walterscheid, R. L. Gravity wave heating and cooling in Saturn's thermosphere. *Eos* **86** (Suppl. 18), abstr. SA24A–06 (2005).
23. Festou, M. C. & Atreya, S. K. Voyager ultraviolet stellar occultation measurements of the composition and thermal profiles of the Saturnian upper atmosphere. *Geophys. Res. Lett.* **9,** 1147–1150 (1982).
24. Atreya, S. K., Waite, J. H., Donahue, T. M., Nagy, A. F. & McConnell, J. C. in *Saturn* (eds Gehrels, T. & Matthews, M. S.) 239–277 (Univ. Arizona Press, Tucson, Arizona, 1984).
25. Smith, G. R. & Hunten, D. M. Study of planetary atmospheres by absorptive occultations. *Rev. Geophys.* **28,** 117–143 (1990).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The simulations described in this study were performed using the HiPerSPACE facility at UCL, funded by the UK Particle Physics and Astronomy Research Council (PPARC). C.G.A.S. acknowledges receipt of a CASE studentship funded by PPARC and Sun Microsystems Ltd.

Author Contributions The thermosphere modelling was carried out by C.G.A.S., A.D.A., G.H.M. and S.M. L.E.M. provided the ionosphere model.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.D.A. (a.aylward@ucl.ac.uk).

Vol 445|25 January 2007|doi:10.1038/nature05513

LETTERS

Comparison of the Hanbury Brown–Twiss effect for bosons and fermions

T. Jeltes[1], J. M. McNamara[1], W. Hogervorst[1], W. Vassen[1], V. Krachmalnicoff[2], M. Schellekens[2], A. Perrin[2], H. Chang[2], D. Boiron[2], A. Aspect[2] & C. I. Westbrook[2]

Fifty years ago, Hanbury Brown and Twiss (HBT) discovered photon bunching in light emitted by a chaotic source[1], highlighting the importance of two-photon correlations[2] and stimulating the development of modern quantum optics[3]. The quantum interpretation of bunching relies on the constructive interference between amplitudes involving two indistinguishable photons, and its additive character is intimately linked to the Bose nature of photons. Advances in atom cooling and detection have led to the observation and full characterization of the atomic analogue of the HBT effect with bosonic atoms[4–6]. By contrast, fermions should reveal an antibunching effect (a tendency to avoid each other). Antibunching of fermions is associated with destructive two-particle interference, and is related to the Pauli principle forbidding more than one identical fermion to occupy the same quantum state. Here we report an experimental comparison of the fermionic and bosonic HBT effects in the same apparatus, using two different isotopes of helium: ^{3}He (a fermion) and ^{4}He (a boson). Ordinary attractive or repulsive interactions between atoms are negligible; therefore, the contrasting bunching and antibunching behaviour that we observe can be fully attributed to the different quantum statistics of each atomic species. Our results show how atom–atom correlation measurements can be used to reveal details in the spatial density[7,8] or momentum correlations[9] in an atomic ensemble. They also enable the direct observation of phase effects linked to the quantum statistics of a many-body system, which may facilitate the study of more exotic situations[10].

Two-particle correlation analysis is an increasingly important method for studying complex quantum phases of ultracold atoms[7–13]. It goes back to the discovery, by Hanbury Brown and Twiss[1], that photons emitted by a chaotic (incoherent) light source tend to be bunched: the joint detection probability is enhanced, compared to that of statistically independent particles, when the two detectors are close together. Although the effect is easily understood in the context of classical wave optics[14], it took some time to find a clear quantum interpretation[3,15]. The explanation relies on interference between the quantum amplitude for two particles, emitted from two source points S_1 and S_2, to be detected at two detection points D_1 and D_2 (see Fig. 1). For bosons, the two amplitudes $\langle D_1|S_1\rangle\langle D_2|S_2\rangle$ and $\langle D_1|S_2\rangle\langle D_2|S_1\rangle$ must be added, which yields a factor of 2 excess in the joint detection probability, if the two amplitudes have the same phase. The sum over all pairs (S_1,S_2) of source points washes out the interference, unless the distance between the detectors is small enough that the phase difference between the amplitudes is less than one radian, or equivalently if the two detectors are separated by a distance less than the coherence length. Study of the joint detection rates versus detector separation along the i direction then reveals a 'bump' whose width l_i is the coherence length along that axis[1,5,16–19]. For a source size s_i (defined as the half width at $e^{-1/2}$ of a gaussian density profile) along the i direction, the bump has a half width at e^{-1} of $l_i = ht/(2\pi m s_i)$, where m is the mass of the particle, t the time of flight from the source to the detector, and h Planck's constant. This formula is the analogue of the formula $l_i = L\lambda/(2\pi s_i)$ for photons, if $\lambda = h/(mv)$ is identified with the de Broglie wavelength for particles travelling at velocity $v = L/t$ from the source to the detector.

For indistinguishable fermions, the two-body wavefunction is antisymmetric, and the two amplitudes must be subtracted, yielding a null probability for joint detection in the same coherence volume. In the language of particles, it means that two fermions cannot have momenta and positions belonging to the same elementary cell of

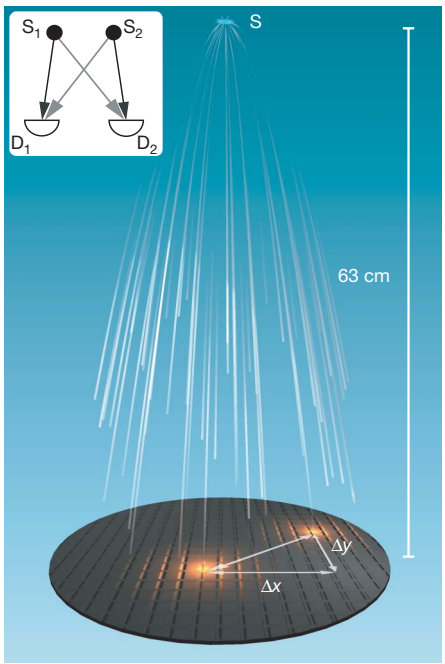


Figure 1 | The experimental set-up. A cold cloud of metastable helium atoms is released at the switch-off of a magnetic trap. The cloud expands and falls under the effect of gravity onto a time-resolved and position-sensitive detector (microchannel plate and delay-line anode) that detects single atoms. The horizontal components of the pair separation Δr are denoted Δx and Δy. The inset shows conceptually the two 2-particle amplitudes (in black or grey) that interfere to give bunching or antibunching: S_1 and S_2 refer to the initial positions of two identical atoms jointly detected at D_1 and D_2.

[1]Laser Centre Vrije Universiteit, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands. [2]Laboratoire Charles Fabry de l'Institut d'Optique, CNRS, Univ. Paris-sud, Campus Polytechnique RD 128, 91127 Palaiseau Cedex, France.

phase space. As a result, for fermions the joint detection rate versus detector separation is expected to exhibit a dip around the null separation. Such a dip for a fermion ensemble must not be confused with the antibunching dip that one can observe with a single particle (boson or fermion) quantum state—for example, resonance fluorescence photons emitted by an individual quantum emitter[20]. In contrast to the HBT effect for bosons, the fermion analogue cannot be interpreted by any classical model, either wave or particle, and extensive efforts have been directed towards an experimental demonstration. Experiments have been performed with electrons in solids[21,22] and in a free beam[23], and with a beam of neutrons[24], but none has allowed a detailed study and a comparison of the pure fermionic and bosonic HBT effects for an ideal gas. A recent experiment using fermions in an optical lattice[25], however, does permit such a study and is closely related to our work.

Here we present an experiment in which we study the fermionic HBT effect for a sample of polarized, metastable ^{3}He atoms (^{3}He*), and we compare it to the bosonic HBT effect for a sample of polarized, but not Bose condensed, metastable ^{4}He atoms (^{4}He*) produced in the same apparatus at the same temperature. We have combined the position- and time-resolved detector, previously used[5,26] for ^{4}He*, with an apparatus with which ultracold samples of ^{3}He* or ^{4}He* have recently been produced[27]. Fermions or bosons at thermal equilibrium in a magnetic trap are released onto the detector, which counts individual atoms (see Fig. 1) with an efficiency of approximately 10%. The detector allows us to construct the normalized correlation function $g^{(2)}(\Delta\mathbf{r})$, that is, the probability of joint detection at two points separated by $\Delta\mathbf{r}$, divided by the product of the single detection probabilities at each point. Statistically independent detection events result in a value of 1 for $g^{(2)}(\Delta\mathbf{r})$. A value larger than 1 indicates bunching, while a value less than 1 is evidence of antibunching.

We produce gases of pure ^{3}He* or pure ^{4}He* by a combination of evaporative and sympathetic cooling in an anisotropic magnetic trap (see Methods). Both isotopes are in pure magnetic substates, with nearly identical magnetic moments and therefore nearly identical trapping potentials, so that trapped non-degenerate and non-interacting samples have the same size at the same temperature. The temperatures of the samples yielding the results of Fig. 2, as measured by the spectrum of flight times to the detector, are 0.53 ± 0.03 μK and 0.52 ± 0.05 μK for ^{3}He* and ^{4}He*, respectively. The uncertainties correspond to the standard deviation of each ensemble. In a single realization, we typically produce 7×10^4 atoms of both ^{3}He* and ^{4}He*. The atom number permits an estimate of the Fermi and Bose–Einstein condensation temperatures of approximately 0.9 μK and 0.4 μK, respectively. Consequently, Fermi pressure in the trapped ^{3}He* sample has a negligible (3%) effect on the trap size, and repulsive interactions in the ^{4}He* sample have a similarly small effect. The trapped samples are therefore approximately gaussian ellipsoids elongated along the x axis with an r.m.s. size of about $110 \times 12 \times 12$ μm^3. To release the atoms, we turn off the current in the trapping coils and atoms fall under the influence of gravity. The detector, placed 63 cm below the trap centre (see Fig. 1), then records the x–y position and arrival time of each detected atom.

The normalized correlation functions $g^{(2)}(0,0,\Delta z)$ along the z (vertical) axis, for ^{3}He* and ^{4}He* gases at the same temperature, are shown in Fig. 2. Each correlation function is obtained by analysing the data from about 1,000 separate clouds for each isotope (see Methods). Results analogous to those of Fig. 2 are obtained for correlation functions along the y axis, but the resolution of the detector in the x–y plane (about 500 μm half width at e^{-1} for pair separation) broadens the signals. Along the x axis (the long axis of the trapped clouds), the expected widths of the HBT structures are one order of magnitude smaller than the resolution of the detector and are therefore not resolved.

Figure 2 shows clearly the contrasting behaviours of bosons and fermions. In both cases we observe a clear departure from statistical independence at small separation. Around zero separation, the fermion signal is lower than unity (antibunching) while the boson signal is higher (bunching). Because the sizes of the ^{3}He* and ^{4}He* clouds at the same temperature are the same, as are the times of flight (pure free fall), the ratio of the correlation lengths is expected to be equal to the inverse of the mass ratio, 4/3. The observed ratio of the correlation lengths along the z axis in the data shown is 1.3 ± 0.2. The individual correlation lengths are also in good agreement with the formula $l_z = ht/(2\pi m s_z)$, where s_z is the source size along z. Owing to the finite resolution, the contrast in the signal, which should ideally go to 0 or 2, is reduced by a factor of order ten. The amount of contrast reduction is slightly different for bosons and fermions, and the ratio should be about 1.5. The measured ratio is 2.4 ± 0.2. This discrepancy has several possible explanations. First, the magnetic field switch-off is not sudden (timescale ~1 ms), and this could affect bosons and fermions differently. Second, systematic errors may be present in our estimate of the resolution function. The resolution, however, does not affect the widths of the observed correlation functions along z, and thus we place the strongest emphasis on this ratio as a test of our understanding of boson and fermion correlations in an ideal gas. More information on uncertainties and systematic errors, as well as a more complete summary of the data, are given in Supplementary Information.

Improved detector resolution would allow a more detailed study of the correlation function, and is thus highly desirable. The effect of the resolution could be circumvented by using a diverging atom lens to demagnify the source[4]. According to the formula $l = ht/(2\pi ms)$, a smaller effective source size gives a larger correlation length. We have tried such a scheme by creating an atomic lens with a blue-detuned, vertically propagating, laser beam, forcing the atoms away from its axis (see Methods). The laser waist was not large compared to the cloud size, and therefore our 'lens' suffered from strong aberrations, but a crude estimate of the demagnification, neglecting aberrations, gives about 2 in the x–y plane. Figure 3 shows a comparison of

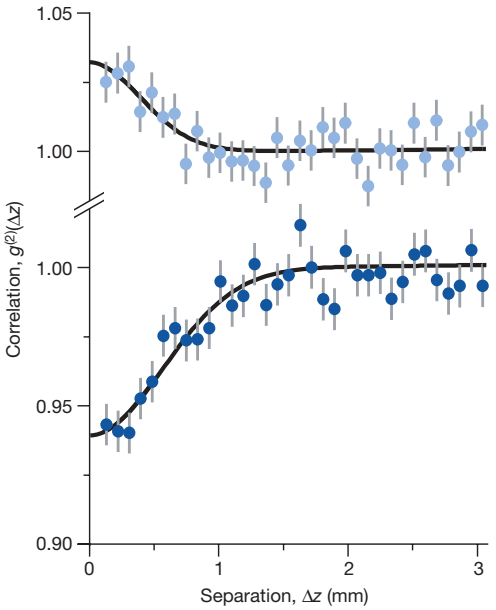


Figure 2 | Normalized correlation functions for ^{4}He* (bosons) in the upper plot, and ^{3}He* (fermions) in the lower plot. Both functions are measured at the same cloud temperature (0.5 μK), and with identical trap parameters. Error bars correspond to the square root of the number of pairs in each bin. The line is a fit to a gaussian function. The bosons show a bunching effect, and the fermions show antibunching. The correlation length for ^{3}He* is expected to be 33% larger than that for ^{4}He* owing to the smaller mass. We find 1/e values for the correlation lengths of 0.75 ± 0.07 mm and 0.56 ± 0.08 mm for fermions and bosons, respectively.

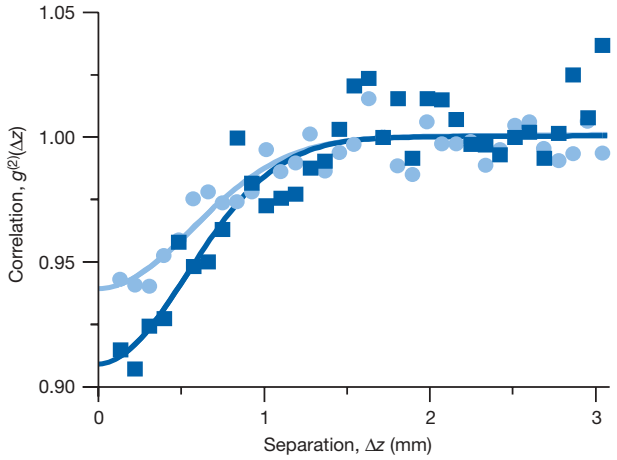


Figure 3 | Effect of demagnifying the source size. We show normalized correlation functions along the z (vertical) axis for ^{3}He*, with (dark blue squares) and without (light blue circles) a diverging atomic lens in the x–y plane. The dip is deeper with the lens, because the increase of the correlation lengths in the x–y plane leads to less reduction of contrast when convolved with the resolution function in that plane.

$g^{(2)}(\Delta z)$ for fermions with and without the defocusing lens. We clearly see a greater antibunching depth, consistent with larger correlation lengths in the x–y plane (we have checked that l_y is indeed increased) and therefore yielding a smaller reduction of the contrast when convolved with the detector resolution function. As expected, the correlation length in the z direction is unaffected by the lens in the x–y plane. Although our atomic lens was far from ideal, the experiment shows that it is possible to modify the HBT signal by optical means.

To conclude, we emphasize that we have used samples of neutral atoms at a moderate density in which interactions do not play any significant role. Care was taken to manipulate bosons and fermions in conditions as similar as possible. Thus the observed differences can be understood as a purely quantum effect associated with the exchange symmetries of wavefunctions of indistinguishable particles.

The possibility of having access to the sign of phase factors in a many-body wavefunction opens fascinating perspectives for the investigation of intriguing analogues of condensed-matter systems, which can now be realized with cold atoms. For instance, one could compare the many-body state of cold fermions and that of 'fermionized' bosons in a one-dimensional sample[28,29]. Our successful manipulation of the HBT signal by interaction with a laser suggests that other lens configurations could allow measurements in position space (by forming an image of the cloud at the detector) or in any combination of momentum and spatial coordinates.

METHODS

Experimental sequence. Clouds of cold ^{4}He* are produced by evaporative cooling of a pure ^{4}He* sample, loaded into a Ioffe–Pritchard magnetic trap[30]. The trapped state is 2^3S_1, $m_J = 1$, and the trap frequency values are 47 Hz and 440 Hz for axial and radial confinement, respectively. The bias field is 0.75 G, corresponding to a frequency of 2.1 MHz for a transition between the $m_J = 1$ and $m_J = 0$ states at the bottom of the trap. After evaporative cooling, we keep the radio frequency evaporation field ('r.f. knife') on at constant frequency for 500 ms, then wait for 100 ms before switching off the trap. In contrast to the experiments of ref. 5, atoms are released in a magnetic-field-sensitive state.

To prepare ^{3}He* clouds, we simultaneously load ^{3}He* and ^{4}He* atoms in the magnetic trap[27]. The trapping state for ^{3}He* is 2^3S_1, $F = 3/2$, $m_F = 3/2$, and axial and radial trap frequencies are 54 Hz and 506 Hz—the difference compared to ^{4}He* is only due to the mass. The two gases are in thermal equilibrium in the trap, so that ^{3}He* is sympathetically cooled with ^{4}He* during the evaporative cooling stage. Once the desired temperature is reached, we selectively eliminate ^{4}He* atoms from the trap using the r.f. knife. The gyromagnetic ratios for ^{4}He* and ^{3}He* are 2 and 4/3 respectively, so that the resonant frequency of the $m = 1$ to $m = 0$ transition for ^{4}He* is 3/2 times larger than the $m = 3/2$ to $m = 1/2$ transition for ^{3}He*. An r.f. ramp from 3 MHz to 1.9 MHz expels all the ^{4}He* atoms from the trap without affecting ^{3}He*. We then use the same trap switch-off procedure to release the ^{3}He* atoms (also in a magnetic-field-sensitive state) onto the detector. We can apply magnetic field gradients to check the degree of spin polarization of either species.

Correlation function. The detailed procedure leading to this correlation is given in ref. 5. Briefly, we convert arrival times to z positions, and then use the three-dimensional positions of each atom to construct a histogram of pair separations $\Delta\mathbf{r}$ in a particular cloud. We then sum the pair distribution histograms for 1,000 successive runs at the same temperature. For separations much larger than the correlation length, this histogram reflects the gaussian spatial distribution of the cloud. To remove this large-scale shape and obtain the normalized correlation function, we divide the histogram by the autoconvolution of the sum of the 1,000 single-particle distributions.

Atom lens experiment. A 300 mW laser beam with an elliptical waist of approximately $100 \times 150\ \mu m^2$ propagates vertically through the trap. The laser frequency is detuned by 300 GHz from the 2^3S_1 to 2^3P_2 transition. After turning off the magnetic trap, and waiting 500 μs for magnetic transients to die away, the defocusing laser is turned on for 500 μs.

Received 15 November; accepted 7 December 2006.

1. Hanbury Brown, R. & Twiss, R. Q. Correlation between photons in two coherent beams of light. *Nature* **177,** 27–29 (1956).
2. Scully, M. O. & Zubairy, M. S. *Quantum Optics* (Cambridge Univ. Press, Cambridge, UK, 1997).
3. Glauber, R. J. in *Quantum Optics and Electronics* (eds DeWitt, C., Blandin, A. & Cohen-Tannoudji, C.) 63–185 (Gordon and Breach, New York, 1965).
4. Yasuda, M. & Shimizu, F. Observation of two-atom correlation of an ultracold neon atomic beam. *Phys. Rev. Lett.* **77,** 3090–3093 (1996).
5. Schellekens, M. *et al.* Hanbury Brown Twiss effect for ultracold quantum gases. *Science* 310, 648–651 (2005); published online 15 September 2005 (doi:10.1126/science.1118024).
6. Öttl, A., Ritter, S., Köhl, M. & Esslinger, T. Correlations and counting statistics on an atom laser. *Phys. Rev. Lett.* **95,** 090404 (2005).
7. Fölling, S. *et al.* Spatial quantum noise interferometry in expanding condensates. *Nature* **434,** 481–484 (2005).
8. Spielman, I. B., Phillips, W. D. & Porto, J. V. The Mott insulator transition in two dimensions. Preprint at ⟨http://arxiv.org/cond-mat/0606216⟩ (2006).
9. Greiner, M., Regal, C. A., Stewart, J. T. & Jin, D. S. Probing pair-correlated fermionic atoms through correlations in atom shot noise. *Phys. Rev. Lett.* **94,** 110401 (2005).
10. Altman, E., Demler, E. & Lukin, M. D. Probing many-body states of ultracold atoms via noise correlations. *Phys. Rev. A* **70,** 013603 (2004).
11. Grondalski, J., Alsing, P. M. & Deutsch, I. H. Spatial correlation diagnostics for atoms in optical lattices. *Opt. Express* **5,** 249–261 (1999).
12. Hellweg, D. *et al.* Measurement of the spatial correlation function of phase fluctuating Bose-Einstein condensates. *Phys. Rev. Lett.* **91,** 010406 (2003).
13. Estève, J. *et al.* Observations of density fluctuations in an elongated Bose gas: ideal gas and quasicondensate regimes. *Phys. Rev. Lett.* **96,** 130403 (2006).
14. Loudon, R. *The Quantum Theory of Light* 3rd edn (Oxford Univ. Press, Oxford, 2000).
15. Fano, U. Quantum theory of interference effects in the mixing of light from phase independent sources. *Am. J. Phys.* **29,** 539–545 (1961).
16. Hanbury Brown, R. & Twiss, R. Q. A test of a new stellar interferometer on Sirius. *Nature* **178,** 1046–1048 (1956).
17. Baym, G. The physics of Hanbury Brown-Twiss intensity interferometry: From stars to nuclear collisions. *Acta Phys. Pol. B* **29,** 1839–1884 (1998).
18. Boal, D. H., Gelbke, C.-K. & Jennings, B. K. Intensity interferometry in subatomic physics. *Rev. Mod. Phys.* **62,** 553–602 (1990).
19. Viana Gomes, J. *et al.* Theory for a Hanbury Brown Twiss experiment with a ballistically expanding cloud of cold atoms. *Phys. Rev. A* **74,** 053607 (2006).
20. Kimble, H. J., Dagenais, M. & Mandel, L. Photon antibunching in resonance fluorescence. *Phys. Rev. Lett.* **39,** 691–695 (1978).
21. Henny, M. *et al.* The fermionic Hanbury Brown and Twiss experiment. *Science* **284,** 296–298 (1999).
22. Oliver, W. D., Kim, J., Liu, R. C. & Yamamoto, Y. Hanbury Brown and Twiss-type experiment with electrons. *Science* **284,** 299–301 (1999).
23. Kiesel, H., Renz, A. & Hasselbach, F. Observation of Hanbury Brown-Twiss anticorrelations for free electrons. *Nature* **418,** 392–394 (2002).
24. Iannuzzi, M., Orecchini, A., Sacchetti, F., Facchi, P. & Pascazio, S. Direct experimental evidence of free-fermion antibunching. *Phys. Rev. Lett.* **96,** 080402 (2006).
25. Rom, T. *et al.* Free fermion antibunching in a degenerate atomic Fermi gas released from an optical lattice. *Nature* **444,** 733–736 (2006).
26. Jagutzki, O. *et al.* A broad-application microchannel-plate detector system for advanced particle or photon detection tasks: Large area imaging, precise multi-hit timing information and high detection rate. *Nucl. Instrum. Methods Phys. Res. A* **477,** 244–249 (2002).

27. McNamara, J. M., Jeltes, T., Tychkov, A. S., Hogervorst, W. & Vassen, W. Degenerate Bose-Fermi mixture of metastable atoms. *Phys. Rev. Lett.* **97,** 080404 (2006).
28. Girardeau, M. Relationship between systems of impenetrable bosons and fermions in one dimension. *J. Math. Phys. (NY)* **1,** 516–523 (1960).
29. Olshanii, M. Atomic scattering in the presence of an external confinement and a gas of impenetrable bosons. *Phys. Rev. Lett.* **81,** 938–941 (1998).
30. Tychkov, A. S. *et al.* Metastable helium Bose-Einstein condensate with a large number of atoms. *Phys. Rev. A* **73,** 031603(R) (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by the access programme of Laserlab Europe. The LCVU group in Amsterdam is supported by the 'Cold Atoms' programme of the Dutch Foundation for Fundamental Research on Matter (FOM) and by the Space Research Organization Netherlands (SRON). The Atom Optics group of LCFIO is a member of the IFRAF institute and of the Fédération LUMAT of the CNRS, and is supported by the French ANR and by the SCALA programme of the European Union.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.I.W. (christoph.westbrook@institutoptique.fr) or W.V. (w.vassen@few.vu.nl).

LETTERS

'Infotaxis' as a strategy for searching without gradients

Massimo Vergassola[1], Emmanuel Villermaux[2] & Boris I. Shraiman[3]

Chemotactic bacteria rely on local concentration gradients to guide them towards the source of a nutrient[1]. Such local cues pointing towards the location of the source are not always available at macroscopic scales because mixing in a flowing medium breaks up regions of high concentration into random and disconnected patches. Thus, animals sensing odours in air or water detect them only intermittently as patches sweep by on the wind or currents[2–6]. A macroscopic searcher must devise a strategy of movement based on sporadic cues and partial information. Here we propose a search algorithm, which we call 'infotaxis', designed to work under such conditions. Any search process can be thought of as acquisition of information on source location; for infotaxis, information plays a role similar to concentration in chemotaxis. The infotaxis strategy locally maximizes the expected rate of information gain. We demonstrate its efficiency using a computational model of odour plume propagation and experimental data on mixing flows[7]. Infotactic trajectories feature 'zigzagging' and 'casting' paths similar to those observed in the flight of moths[8]. The proposed search algorithm is relevant to the design of olfactory robots[9–11], but the general idea of infotaxis can be applied more broadly in the context of searching with sparse information.

Chemotactic search strategies based on local concentration gradients require the concentration to be high enough to ensure that its average difference measured at two nearby locations is larger than typical fluctuations[1,12] (see also Supplementary Information). The signal-to-noise ratio depends of course on the averaging time and might be improved by waiting. However, the average concentration may be decaying rapidly (for example, exponentially) with distance away from the source, and in this weak signal-to-noise (dilute) case the waiting time becomes huge. An example of organisms performing olfactory search in a dilute limit is provided by moths which use pheromones to locate their mates[2–6]. Moths are known to proceed upwind by way of counterturning patterns of extended ('casting') or limited ('zigzagging') crosswind width, thought to correlate with low and high rates of odour detection. A practical situation involving the challenge of searching in dilute conditions is encountered in the design of 'sniffers'[9–11]—robots that track chemicals emitted by drugs, chemical leaks, explosives and mines. Existing methods apply to high-concentration conditions, where chemotactic[13–16] or plume-tracking strategies[17–21] might be used.

In the dilute limit, the searcher detects odour in a sporadic sequence of distinct events arising from its encounters with patches of fluid (or air) where turbulent mixing has failed to dissipate the advected odour down to a level below the detectability threshold[22–24]. These detection events, or 'hits', are separated by wide 'voids' with no detectible signal. Because the probability of odour encounter depends on the distance from the source, the set of encounters that occurred at times $\{t_i\}$ along the search trajectory $\mathbf{r}(t)$ carries information about the source location. We shall use $\mathcal{T}_t$ to denote times and coordinates of these hits.

In the spirit of coding theory, the trace $\mathcal{T}_t$ might be thought of as a message, sent by the source and transmitted to the searcher with strong noise due to the random nature of odour propagation in the turbulent medium. Decoding of the message is implemented using Bayes' formula to construct, given the received signal, the posterior probability distribution $P_t(\mathbf{r}_0)$ for the unknown location of the source $\mathbf{r}_0$ (see Methods, and similar independent arguments in ref. 25). The subscript t reminds us that $\mathcal{T}_t$ and $P_t(\mathbf{r}_0)$ are dynamical objects, continuously updated with time. The specific decoding protocol depends of course on the nature of the detection events and the transmitting medium. For concreteness, we treat here two cases: (1) experimental time-course data for a mixing flow; and (2) a model where detectable 'particles' (which represent patches of detectable odours) are emitted by the source at rate R, have a finite lifetime τ, propagate with effective diffusivity D and are advected by a mean current or wind $\mathbf{V}$. The decoding protocol requires knowing the probability of odour encounters as a function of the distance to the source. This function can be computed analytically for model (2) or estimated from experimental data for case (1), as detailed in Supplementary Information. The latter method is quite general and might be applied to other cases as well.

Given a probability distribution $P(\mathbf{r}_0)$ for the location of the source, we can show (see Supplementary Information) that the expected search time $\langle T \rangle$ is bounded by $\langle T \rangle \geq e^{S-1}$, where S is Shannon's entropy for the distribution $S \equiv -\int d\mathbf{x} P(\mathbf{x}) \ln P(\mathbf{x})$ (refs 26, 27). The latter quantifies how spread-out the distribution is, and goes to zero when the position of the source is localized to one site, that is, is known. The rate of acquisition of information is quantified by the rate of reduction of entropy[26,27] (see also Supplementary Information). The main problem for the searcher is that the real probability distribution is unknown (to it) and must be estimated from the available data: the history of its odour encounters. As information accumulates, the entropy of the estimated distribution decreases and with it the expected time to locate the source. The searcher is faced with conflicting choices of either proceeding with its current information (that is, going to the estimated most probable source location), or alternatively, pausing to gather more information and obtain a more reliable estimate of the source distribution. The problem of dealing with only partially reliable information is quite general, and has received a systematic formulation in learning theory in terms of the 'exploration versus exploitation trade-off' to be struck for effective learning[28]. In the search context, 'exploitation' of the currently estimated $P_t(\mathbf{r}_0)$ by chasing locations of maximal estimated probability is very risky, because it can lead off the track. The most conservative 'exploration' approach is to accumulate information before taking any step. This strategy is safe but not productive,

[1]CNRS URA 2171, Institut Pasteur, "In Silico Genetics", 75724 Paris Cedex 15, France. [2]Université Aix Marseille 1, IRPHE, Technopole Chateau Gombert F-13384 Marseille, France. [3]Kavli Institute for Theoretical Physics, University of California, Santa Barbara, California 93106, USA.

and is inferior to more active exploration—for example, systematic search in a particular sector[29].

To balance exploration and exploitation, we propose the following 'infotaxis' strategy. At each time step, the searcher chooses the direction that locally maximizes the expected rate of information acquisition. Specifically, the searcher chooses, among the neighbouring sites on a lattice and standing still, the move that maximizes the expected reduction in entropy of the posterior probability field.

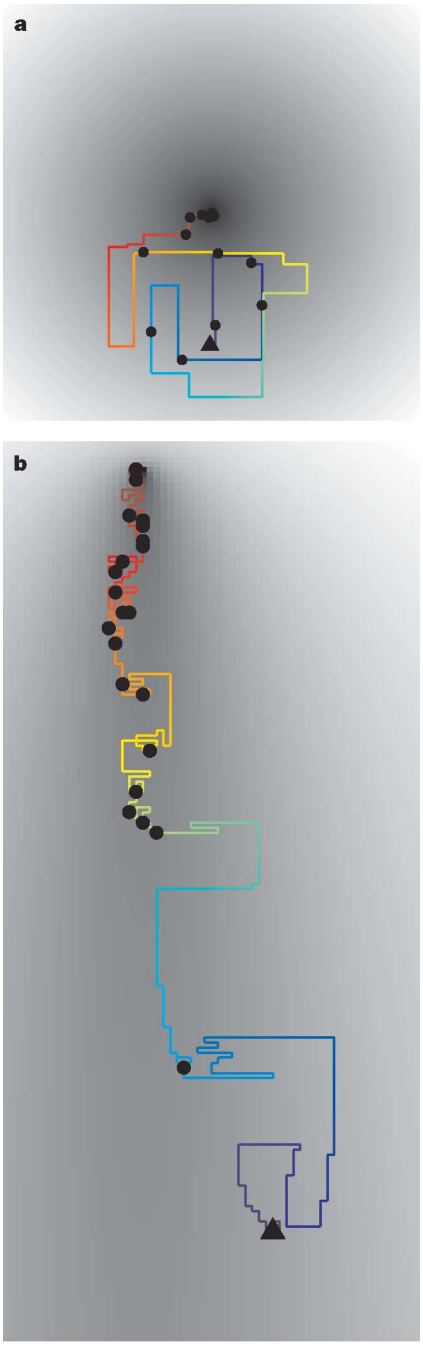


Figure 1 | Typical infotactic trajectories. **a**, Without wind; **b**, with wind. Simulations are performed for a model of odour spreading where detectable 'particles' are emitted at rate R, have a lifetime τ, propagate with diffusivity D (combining turbulent and molecular diffusion) and are advected by a mean wind $\mathbf{V}$. The wind in **b** is directed downwards. The greyscale represents the mean detection rate, decaying exponentially at large distances. In each panel, the searcher starts from the black filled triangle, the colour code on the trajectories is linear in the elapsed time, and odour detections are indicated by black filled circles. Note the long lags with no particle detections, characteristic of searches in dilute conditions.

Expectations are based on the information currently available, that is, the field $P_t(\mathbf{r}_0)$ itself. The intuitive idea is that entropy decreases (and thus information accumulates) faster close to the source because cues arrive at a higher rate, hence tracking the maximum rate of information acquisition will guide the searcher to the source much like concentration gradients in chemotaxis.

Suppose that the searcher has arrived at $\mathbf{r}$ at time t, and gathered information is stored into the field $P_t(\mathbf{r}_0)$ having entropy S. The variation of entropy expected upon moving to one of the neighbouring points $\mathbf{r}_j$ (or standing still) is:

$$\overline{\Delta S}(\mathbf{r}\mapsto\mathbf{r}_j) = P_t(\mathbf{r}_j)[-S] + \left[1-P_t(\mathbf{r}_j)\right]\left[\rho_0(\mathbf{r}_j)\Delta S_0 + \rho_1(\mathbf{r}_j)\Delta S_1 + \ldots\right] \quad (1)$$

The first term on the right-hand side corresponds to finding the source, that is, P_{t+1} becoming a δ-function and entropy becoming zero, which occurs with estimated probability $P_t(\mathbf{r}_j)$. The second term on the right-hand side corresponds to the alternative case when the source is not found at $\mathbf{r}_j$. Symbols $\rho_k(\mathbf{r}_j)$ denote the probability that k detections be made at $\mathbf{r}_j$ during a time-step Δt, given by a Poisson law $\rho_k = h^k e^{-h}/k!$ for independent detections. The expected number of hits is estimated as $h(\mathbf{r}_j) \equiv \Delta t \int P_t(\mathbf{r}_0) R(\mathbf{r}_j|\mathbf{r}_0)\, d\mathbf{r}_0$, with $R(\mathbf{r}|\mathbf{r}_0)$ denoting the mean rate of hits at position $\mathbf{r}$ if the source is located in $\mathbf{r}_0$ (see Methods). The symbols ΔS_k in equation (1) denote the change of entropy between the fields $P_{t+1}(\mathbf{r}_0)$ and $P_t(\mathbf{r}_0)$. Two effects contribute to ΔS_k: first, $P_{t+1}(\mathbf{r}_j) \equiv 0$ because the source was not found; and second, the estimated posterior probabilities are modified by the k cues received. The first term on the right-hand side of equation (1) is the exploitative term, weighing only the event that the source is found at the point $\mathbf{r}_j$ and favouring motion to maximum likelihood points. The second term on the right-hand side of equation (1) is the information gain from receiving additional cues. It appears even when the searcher does not move, and thus represents conservative 'exploration'. Thus we explicitly see that infotaxis naturally combines exploitative and exploratory tendencies (see Supplementary Information for details of this point and for quantitative comparisons among different strategies).

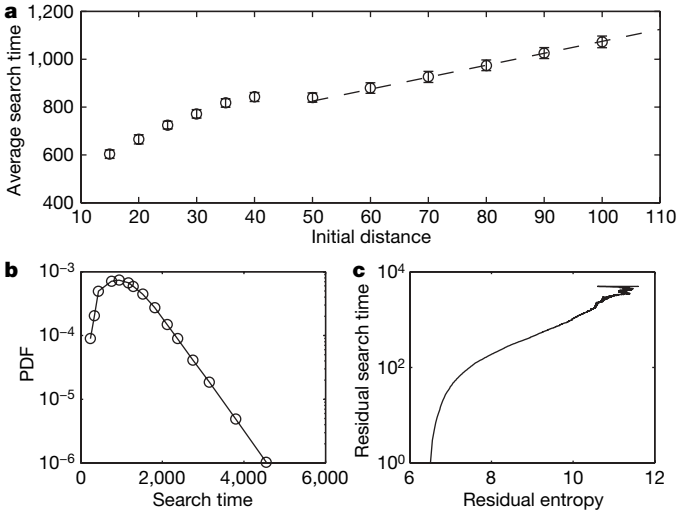


Figure 2 | Quantitative characterization of infotaxis searches. **a**, Scaling of the average search time with the initial distance to the source. The mean path of particles during their lifetime is 50. The linear scaling at large distances compares favourably with the exponential time needed to average out concentration noise. Error bars indicate s.d. **b**, The exponential decay of the search time probability distribution function (PDF), indicating that infotaxis is not plagued by strong fluctuations. **c**, The residual time to locate the source plotted against the entropy of the estimated source location PDF. The exponential dependence of the residual time indicates that reducing entropy is effective in ensuring a rapid search process.

Infotactic paths are illustrated in Fig. 1, which presents the result of a numerical simulation using the model of odour propagation described in Methods. In the absence of wind (Fig. 1a), $P_t(\mathbf{r})$ is rotationally symmetric around the starting point and the searcher starts spiralling around it (as is observed with sea urchin sperm[30]). Interestingly, in the absence of hits the radius of the spiral increases in a scale invariant manner, making an approximately Archimedean spiral. As time progresses and information is gathered along the trajectory, the direction towards the source emerges as the preferential one, finally leading the searcher to the source. In the presence of wind (Fig. 1b), the search alternates phases of consistent progression upwind with phases of wider crosswind excursion and even downwind movements similar to the classical casting and zigzagging patterns observed during bird and moth flights[8].

To quantify the performance of the proposed search algorithm, we examine (Fig. 2a) the scaling of the average search time with the initial distance to the source for the model without wind (the most difficult one). The linear dependence found for large initial distances should be contrasted with the exponential dependence needed to average out the concentration noise (see Supplementary Information). Furthermore, random walking searchers would often attain the boundaries of the box where the search is taking place. The corresponding probability distribution of search times would have a long tail, decaying as $1/T^2$, contrary to the exponential decay shown in Fig. 2b. Figure 2c shows the relation between search times and the entropy of the posterior field $P_t(\boldsymbol{r}_0)$, which supports the theoretical bound mentioned above.

Figure 3 presents an infotactic path generated in a simulation using experimental measurements of dye concentration in a turbulent flow[7]. Hits occur when the searcher encounters concentration above a threshold, which we chose sufficiently high to keep the number of hits low. Simulations indicate that the infotactic strategy is robust with respect to the searcher's model of the turbulent medium and to fluctuations and inhomogeneities of the medium. Indeed, even the simplistic hypothesis of time-independent odour encounters does not hinder the search. Modelling of the turbulent medium might be further improved by accounting for temporal correlations of odour plume encounter and velocity fluctuations (see Supplementary Information).

We have presented an olfactory search algorithm that works in the dilute limit corresponding to weak sources in realistic flows. These are the conditions encountered in applications of olfactory robots and by various living creatures. When comparing our results with the behaviour of living creatures, we need to bear in mind that similarity in trajectories does not imply identity of causal mechanisms and decision processes[10,15]. Still, it is worth remarking that the olfactory search motion observed with moths and birds exhibits a pattern of extended crosswind casts and zigzags similar to the one generated in Fig. 1b by the general principle of local maximization of information gain. We note that the dilute limit also describes the case of 'odour' diffusion at the molecular scale provided that the searcher can detect single molecules. This situation may apply, for example, in the case of sea urchin sperm, which responds to a single molecule of resact, a chemoattractant peptide[30]. It is difficult to imagine a single cell

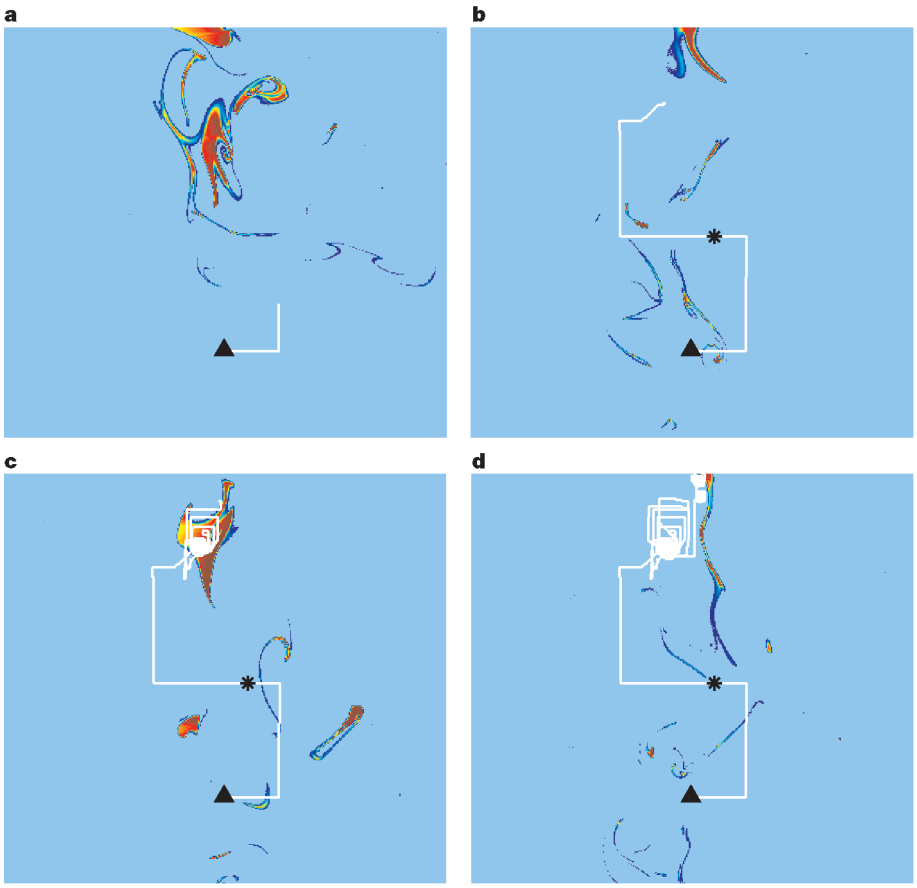


Figure 3 | Simulation of infotaxis using mixing flow experimental data[7]. Snapshots (false colours; red corresponds to high concentrations) of dye concentration levels are acquired from mixing flow experimental data, and the trajectory of the searcher is numerically simulated in the resulting sequence of fields. Snapshots and trajectories are shown at four successive times in panels **a**–**d**. Light blue regions correspond to concentrations below the detection threshold. The black star denotes an isolated odour detection event. The infotactic strategy is robust with respect to fluctuations and inhomogeneities of the medium (for example, of the wind direction) and its modelling by the searcher. Robustness stems from the tracking of information rather than estimated maximum likelihood locations.

performing complex computations like those required for infotaxis, as described above. It will be interesting to explore heuristic approximations simple enough to be plausible for single cells. Finally, we note that the general information-theoretic approach described above in the olfactory search context applies more broadly to any situation where competing demands of exploration and exploitation must be efficiently balanced.

METHODS

Estimation of the posterior probability distribution. The probability distribution posterior to experiencing a trace $\mathcal{T}_t$ of uncorrelated odour encounters is given by:

$$P_t(\mathbf{r}_0) = \frac{\mathcal{L}_{\mathbf{r}_0}(\mathcal{T}_t)}{\int \mathcal{L}_{\mathbf{x}}(\mathcal{T}_t)\mathrm{d}\mathbf{x}} = \frac{\exp\left[-\int_0^t R(\mathbf{r}(t')|\mathbf{r}_0)dt'\right]\prod_{i=1}^{H} R(\mathbf{r}(t_i)|\mathbf{r}_0)}{\int \exp\left[-\int_0^t R(\mathbf{r}(t')|\mathbf{x})dt'\right]\prod_{i=1}^{H} R(\mathbf{r}(t_i)|\mathbf{x})\mathrm{d}\boldsymbol{x}} \quad (2)$$

Here, H is the number of hits along the trajectory, the t_is are the corresponding times and $\mathcal{L}_{\mathbf{r}_0}(\mathcal{T}_t)$ is the likelihood of observing the trace $\mathcal{T}_t$ of odour encounters for a source located at $\mathbf{r}_0$. This expression is supplemented by the prescription that visited regions where the source was not found have zero probability. Note that $P_{t+\Delta t}(\mathbf{r}_0)$ factorizes as $P_t(\mathbf{r}_0)$ times a term that depends on the hits received in the Δt interval. Thus, keeping track of the whole trajectory and the history of detections is not required. The expression for $P_t(\mathbf{r}_0)$ is derived by taking the mean 'hit' rate during an infinitesimal interval $\mathrm{d}t$ to be $R(\mathbf{r}|\mathbf{r}_0)\mathrm{d}t$ and the number of hits to be Poisson distributed. The function $R(\mathbf{r}|\mathbf{r}_0)$ appearing in equation (2) denotes the mean rate of hit encounters at position $\mathbf{r}$ for a source located at $\mathbf{r}_0$. For the model where detectable 'particles' are emitted by the source at rate R, have a finite lifetime τ, propagate with isotropic effective diffusivity D (which parameterizes the combined effect of turbulent and molecular diffusion) and are (possibly) advected by a mean current or wind $\mathbf{V}$, the function can be computed analytically, as described in Supplementary Information. The result for the three dimensional case is:

$$R(\mathbf{r}|\mathbf{r}_0) = \frac{aR}{|\mathbf{r}-\mathbf{r}_0|} e^{-\frac{|\mathbf{r}-\mathbf{r}_0|}{\lambda}} e^{-\frac{(y-y_0)V}{2D}}; \quad \lambda = \sqrt{\frac{D\tau}{1+\frac{V^2\tau}{4D}}} \quad (3)$$

where a is the size of the searcher, and the coordinates are chosen to have the wind blowing along the y axis in the negative direction. A similar expression holds in two dimensions. In both cases, the rate function decreases exponentially at large distances (anisotropically in the presence of wind). For experimental data on mixing flows, the parameters in the rate function are estimated from the data as detailed in Supplementary Information.

Mixing flow experiment and simulation parameters. Parameters used in the simulations are as follows. Figs 1a and 2: emission rate of the source $R = 1$, particle lifetime $\tau = 2{,}500$ and diffusivity $D = 1$. The typical travel distance during a lifetime is approximately $\sqrt{D\tau} = 50$. The search space is a grid 512×512, and equation (1) was evaluated for five possible actions at each time step (moves to the four neighbours and standing still). Figure 1b: wind blows downwards with unit speed, the emission rate $R = 1/2$ and the lifetime of particles $\tau = 150$, the initial vertical separation to the source. Figure 3: experimental data are generated injecting dye by a 8 mm tube in the far field of a jet, a large-scale (integral scale $\sim$10 cm) sustained turbulent flow. The pictures in the figure show the random advection of the dye downstream of the injection point (for about three integral scales). Dye is passively transported by the flow, as odours are. Odour detection events correspond to encounters with dye concentrations above a threshold fixed at about five times the average intensity level. The mean velocity (pointing downward in Fig. 3) is about 4 cm s^{-1} (its measurement by olfactory robots might be realized using standard anemometers), and the root-mean-square velocity is about 30% of the mean velocity, giving a Reynolds number of $\sim 10^3$. Snapshots of the field are acquired at a frequency of 200 Hz. Coherent odour patches make the searcher spiral around the location of the encounters due to correlations among the detections. Yet, the resulting search process is just twice as long as for a model with independent hits, and might be further accelerated by accounting for time-correlations, as detailed in Supplementary Information.

Received 1 May; accepted 14 November 2006.

1. Berg, B. C. *Random Walks in Biology* (Princeton Univ. Press, Princeton, 1993).
2. Payne, T. L., Birch, M. C. & Kennedy, M. C. (eds) *Mechanisms in Insect Olfaction* (Clarendon, Oxford, 1986).
3. Murlis, J., Elkinton, J. S. & Cardé, R. T. Odor plumes and how insects use them. *Annu. Rev. Entomol.* **37,** 505–532 (1992).
4. Dusenbery, D. B. *Sensory Ecology: How Organisms Acquire and Respond to Information* (Freeman, New York, 1992).
5. Mafra-Neto, A. & Cardé, R. T. Fine-scale structure of pheromone plumes modulates upwind orientation of flying moths. *Nature* **369,** 142–144 (1994).
6. Hansson, B. S. (ed.) *Insect Olfaction* (Springer, Berlin, 1999).
7. Villermaux, E. & Duplat, J. Mixing as an aggregation process. *Phys. Rev. Lett.* **91,** 184501 (2003).
8. Kennedy, J. S. Zigzagging and casting as a preprogrammed response to wind-borne odour: A review. *Physiol. Entomol.* **27,** 58–66 (1983).
9. Russell, R. A. *Odor Detection by Mobile Robots* (World Scientific, Singapore, 1999).
10. Webb, B. Robots in invertebrate neuroscience. *Nature* **417,** 359–363 (2002).
11. Marques, L. & de Almeida, A. (eds) Special issue on mobile robots olfaction. *Auton. Robots* **20,** 183–287 (2006).
12. Berg, H. C. & Purcell, E. M. Physics of chemoreception. *Biophys. J.* **20,** 193–219 (1977).
13. Ishida, H., Kagawa, Y., Nakamoto, T. & Moriizumi, T. Odor-source localization in the clean room by an autonomous mobile sensing system. *Sens. Actuators B* **33,** 115–121 (1996).
14. Kuwana, Y., Nagasawa, S., Shimoyama, I. & Kanzaki, R. Synthesis of the pheromone oriented behaviour of silkworm moths by a mobile robot with moth antennae as pheromone sensors. *Biosens. Bioelectron.* **14,** 195–202 (1999).
15. Grasso, F. W., Consi, T. R., Mountain, D. C. & Atema, J. Biomimetic robot lobster performs chemo-orientation in turbulence using a pair of spatially separated sensors: Progress and challenges. *Rob. Auton. Syst.* **30,** 115–131 (2000).
16. Russell, R. A., Bab-Hadiashar, A., Shepherd, R. L. & Wallace, G. G. A comparison of reactive robot chemotaxis algorithms. *Rob. Auton. Syst.* **45,** 83–97 (2003).
17. Belanger, J. H. & Arbas, E. Behavioral strategies underlying pheromone-modulated flights in moths: Lessons from simulation studies. *J. Comp. Physiol. A* **183,** 345–360 (1998).
18. Li, W., Farrell, J. A. & Cardé, R. T. Tracking of fluid-advected odor plumes: strategies inspired by insect orientation to pheromone. *Adapt. Behav.* **9,** 143–170 (2001).
19. Farrell, J. A., Pang, S. & Li, W. Plume mapping via hidden Markov methods. *IEEE Trans. Syst. Man Cybern. B* **33,** 850–863 (2003).
20. Farrell, J. A., Pang, S. & Li, W. Chemical plume tracing via an autonomous underwater vehicle. *IEEE J. Ocean. Eng.* **30,** 428–442 (2005).
21. Ishida, H., Nakayama, G., Nakamoto, T. & Moriizumi, T. Controlling a gas/odor plume-tracking robot based on transient responses of gas sensors. *IEEE Sensors J.* **5,** 537–545 (2005).
22. Murlis, J. & Jones, C. D. Fine-scale structure of odor plumes in relation to insect orientation to distant pheromone and other attractant sources. *Physiol. Entomol.* **6,** 71–86 (1981).
23. Shraiman, B. I. & Siggia, E. D. Scalar turbulence. *Nature* **405,** 639–646 (2000).
24. Falkovich, G., Gawędzki, K. & Vergassola, M. Particles and fields in fluid turbulence. *Rev. Mod. Phys.* **73,** 913–975 (2001).
25. Pang, S. & Farrell, J. A. Chemical plume source localization. *IEEE Trans. Syst. Man Cybern. B* (in the press).
26. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27,** 379–423 623–656 (1948).
27. Cover, T. M. & Thomas, J. A. *Elements of Information Theory* (Series in Telecommunication, Wiley, New York, 1991).
28. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, Massachusetts, 1998).
29. Balkovsky, E. & Shraiman, B. I. Olfactory search at high Reynolds number. *Proc. Natl Acad. Sci. USA* **99,** 12589–12593 (2002).
30. Kaupp, U. B. *et al.* The signal flow and motor response controlling chemotaxis of sea urchin sperm. *Nature Cell Biol.* **5,** 109–117 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was done during the visits of M.V. and E.V. to KITP, and was supported by the ARO. E.V. is a member of the Institut Universitaire de France.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to B.I.S. (shraiman@kitp.ucsb.edu).

LETTERS

Transformation of spin information into large electrical signals using carbon nanotubes

Luis E. Hueso[1]†, José M. Pruneda[2,3]†, Valeria Ferrari[4]†, Gavin Burnell[1]†, José P. Valdés-Herrera[1,5], Benjamin D. Simons[4], Peter B. Littlewood[4], Emilio Artacho[2], Albert Fert[6] & Neil D. Mathur[1]

Spin electronics (spintronics) exploits the magnetic nature of electrons, and this principle is commercially applied in, for example, the spin valves of disk-drive read heads. There is currently widespread interest in developing new types of spintronic devices based on industrially relevant semiconductors, in which a spin-polarized current flows through a lateral channel between a spin-polarized source and drain[1,2]. However, the transformation of spin information into large electrical signals is limited by spin relaxation, so that the magnetoresistive signals are below 1% (ref. 2). Here we report large magnetoresistance effects (61% at 5 K), which correspond to large output signals (65 mV), in devices where the non-magnetic channel is a multiwall carbon nanotube that spans a 1.5 μm gap between epitaxial electrodes of the highly spin polarized[3,4] manganite $La_{0.7}Sr_{0.3}MnO_3$. This spintronic system combines a number of favourable properties that enable this performance; the long spin lifetime in nanotubes due to the small spin–orbit coupling of carbon; the high Fermi velocity in nanotubes that limits the carrier dwell time; the high spin polarization in the manganite electrodes, which remains high right up to the manganite–nanotube interface; and the resistance of the interfacial barrier for spin injection. We support these conclusions regarding the interface using density functional theory calculations. The success of our experiments with such chemically and geometrically different materials should inspire new avenues in materials selection for future spintronics applications.

We show how carbon nanotubes (CNTs) can solve a long-standing spintronics challenge—namely, the injection of spins into a non-magnetic material and the subsequent transformation of the spin information into a large electrical signal. This challenge began in 1990 with the introduction[5] of the spin-transistor concept. The idea is to use a gate voltage to manipulate spins injected into a semiconductor channel between ferromagnetic contacts. In all spin-transistor concepts based on similar structures[1,2], the prerequisite is a significant magnetoresistance ($MR = \Delta R/R_P$) of the order of unity or larger, where $\Delta R = R_{AP} - R_P$ is the resistance change when a magnetic field alters the relative orientation of the magnetizations of source and drain electrodes between antiparallel (AP) and parallel (P). Experimental MR values[2] have been limited to ~0.1–1%. Here we show why replacing the semiconductor channel with a CNT permits a value of MR = 61%, and thus a significant voltage change of 65 mV.

CNTs are robust, easy to manipulate, and have been successfully used[6] in proof-of-principle field-effect transistors, quantum dots and logic gates. For spintronics, the weak spin–orbit coupling permits a long spin lifetime. Here we also exploit the large[7] CNT Fermi velocity v_F, related to the zero bandgap character of the electronic structure and the resulting linear dispersion[6]. However, it is far from obvious whether spin information can survive long-distance transport, given the likelihood of defects and contamination.

Our study of CNTs with ferromagnetic electrodes represents a fusion of molecular[8] and spin electronics[1], that is, molecular spintronics. In this nascent field, MR effects are typically confined to low temperatures in devices based on octanethiol[9], C_{60} (ref. 10) or CNTs[11–14]. These CNT devices used electrodes made of cobalt[11,14], Pd-Ni (ref. 12) or GaMnAs (ref. 13), and MR effects were studied at low biases and temperatures. The MR is generally small (~10%), and inversions of sign, either from sample to sample, or as a function of voltage[11–14], are related to Coulomb blockade and level quantization. We avoid these effects by measuring MR up to 120 K, and under biases exceeding 25 mV. This voltage is sufficient, given that the Coulomb blockade energy[7] for similar CNTs with albeit different contacts is ~0.1 meV, and given also a level spacing of $hv_F/2L \approx 0.8$ meV for an undoped metallic tube of length $L = 2\,\mu m$ with $v_F = 0.8 \times 10^6\,m\,s^{-1}$ (ref. 7). High-bias MR measurements are possible because naturally occurring tunnel barriers at each electrode–CNT interface limit the current and thus unwanted heating, and significant because unlike MR values alone they represent large output signals.

In this Letter, we present devices (Fig. 1 and Methods) in which epitaxial electrodes of the pseudo-cubic perovskite manganite $La_{0.7}Sr_{0.3}MnO_3$ (LSMO) are electrically connected by a single multiwall CNT, which lies on top of the electrodes—in contrast to standard nanotube device geometries[6]. At low temperatures, the conduction in LSMO exhibits a very high spin polarization[3,4] approaching 100%, whereas the figure for elemental ferromagnets[15] is <40%. Moreover, as LSMO is an oxide, it displays environmental stability, so molecules may be introduced *ex situ*. However, it is not *a priori* known whether spin information can be efficiently transmitted between two materials that possess very different geometries and chemistries.

Similar and reproducible zero-field current–voltage (I–V) characteristics (Supplementary Fig. S1) were seen in 12 devices. Four of these show the large MR effects discussed later, and the other eight show no MR effects. Our CNT–LSMO interfaces behave like tunnel junctions in two respects: first, the $I(V)$ curves are strongly nonlinear; and second, the low-bias (25 mV), low-temperature (5 K) resistance V/I = 10–100 MΩ of our 12 devices is 3–4 orders of magnitude larger than the inverse of the quantum conductance e^2/h typically seen[12,14,16]

[1]Department of Materials Science, University of Cambridge, Pembroke Street, Cambridge CB2 3QZ, UK. [2]Department of Earth Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EQ, UK. [3]Institut de Ciencia de Materials de Barcelona, CSIC Campus UAB, 08193 Bellaterra, Barcelona, Spain. [4]Cavendish Laboratory, University of Cambridge, JJ Thomson Avenue, Cambridge CB3 0HE, UK. [5]Nanoscience Centre, University of Cambridge, JJ Thomson Avenue, Cambridge CB3 0FF, UK. [6]Unité Mixte de Physique CNRS-Thales, TRT, 91767 Palaiseau and Université Paris-Sud, 91405 Orsay, France. †Present addresses: ISMN-CNR, via Gobetti 101, 40129 Bologna, Italy (L.E.H.); Department of Physics, University of California, Berkeley, California 94720, USA (J.M.P.); Departamento de Física, Comisión Nacional de Energia Atómica, Gral. Paz 1499, 1650 San Martín, Buenos Aires, Argentina (V.F.); School of Physics and Astronomy, University of Leeds, Leeds LS2 9JT, UK (G.B.).

for nanotubes between standard metallic electrodes (~13 kΩ). Note that tunnel barriers are generally found at the interfaces between LSMO and metals[17]. However, the interfacial resistance of our devices[18] is not unduly high, and falls within the wide range of values[17] associated with metal–LSMO contacts.

The observed tunnel barriers may be understood through first-principles calculations (Methods) of the electronic structure of an LSMO–CNT interface. The CNT is not significantly altered when contacted by LSMO (Fig. 2a), suggesting that the barrier is localized at the interface, and that our experiments may be insensitive to CNT type and orientation. The Kohn–Sham potential[19]—the simplest estimate of the local energy of a tunnelling electron—shows a barrier (Fig. 2 inset) whose height somewhat exceeds the characteristic CNT kinetic energy (as estimated by the inverse density of states). This is a prerequisite for a tunnel barrier, although the ratio of height to kinetic energy suggests a decay length not much smaller than the barrier width itself, and therefore a relatively high transmission probability. Note that our first-principles calculations also help explain the large MR, because they indicate (Fig. 2b) that the LSMO surface is highly spin polarized despite a pronounced interfacial state ~0.2 eV below the Fermi level.

Our main result is the observation of a large device MR value of 61% (Fig. 3) that arises because of sharp and irreversible switching of the LSMO electrode magnetizations between parallel and antiparallel. Three other working devices showed values of 54%, 72% and 53% (Supplementary Information). These four MR values are much higher than the values generally observed[11–14] with CNTs between other ferromagnetic contacts (~10%). We now discuss why the use of a CNT in place of a standard semiconductor permits the large MR.

The MR of a structure composed of a conduction channel connected to a ferromagnetic source and drain through spin-dependent interface resistances (for example, a tunnel junction) can be expressed[20,21] as:

$$\mathrm{MR} = \frac{\Delta R}{R_\mathrm{P}} \equiv \frac{R_\mathrm{AP} - R_\mathrm{P}}{R_\mathrm{P}} \equiv \frac{\gamma^2/(1-\gamma^2)}{1+\tau_\mathrm{n}/\tau_\mathrm{sf}} \quad (1)$$

where γ is the electrode spin polarization, or more formally the interfacial spin-asymmetry coefficient that influences the spin-dependent interface resistance $r_{\uparrow(\downarrow)} = 2(1 \mp \gamma)\, r_\mathrm{b}^*$ where r_b^* is the mean value of the spin-independent interface resistance, τ_sf is the spin lifetime and τ_n is the dwell time of the electrons in the channel:

$$\tau_\mathrm{n} = 2\,L/(v_\mathrm{N}\,\bar{t}_\mathrm{r}) \quad (2)$$

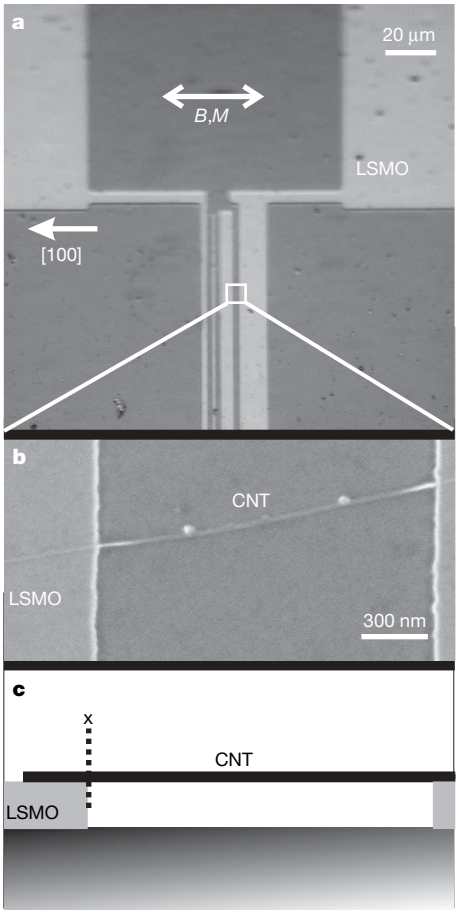


Figure 1 | LSMO–CNT–LSMO device. **a**, Optical micrograph of four variable-width LSMO electrodes, and two of the four associated contact pads. In electrically conducting devices, two adjacent electrodes were connected by an overlying CNT, in regions such as that in the white square. Magnetic fields B were applied along the orthorhombic [100] direction in which the magnetization M is expected to lie due to uniaxial magnetocrystalline anisotropy. **b**, Scanning electron microscope image of a CNT running between LSMO electrodes; magnified view corresponding to the boxed area in **a**. **c**, Schematic side view of **b** with the plane through the CNT at the edge of the LSMO electrode denoted ×.

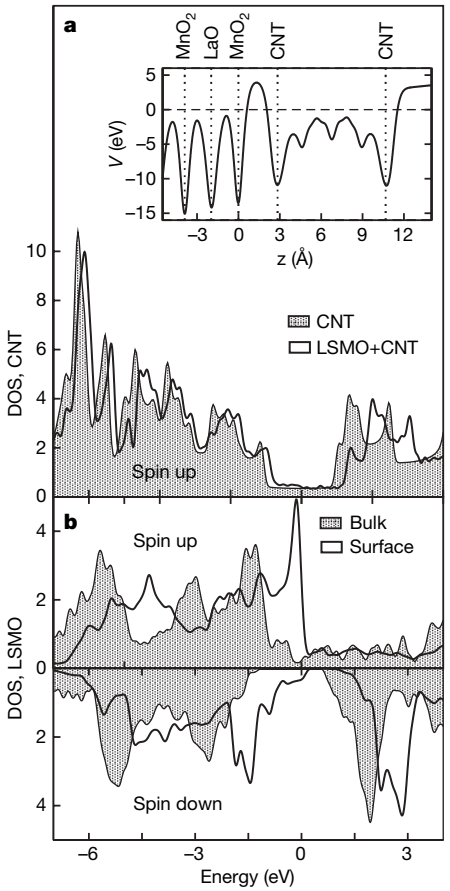


Figure 2 | First-principles calculations of device interfaces. Projected density of states (DOS) on **a**, the basis functions of an isolated CNT (shaded), and a CNT lying on LSMO (unshaded). **b**, the projected DOS onto the first MnO_2+(La,Sr)O layer of the LSMO slab (unshaded) and onto bulk LSMO (shaded). Fermi levels are aligned at zero energy, and only up spins are shown in **a** as up–down differences in the CNT DOS are barely visible at this scale (there is a net spin polarization of +0.01 electrons $Å^{-1}$). Inset, the Kohn–Sham potential seen by electrons in the vicinity of the LSMO–CNT interface. It has been integrated for each value of z (normal to the LSMO surface) in the rectangle defined by the projection of the CNT onto the x–y plane. The origin of potential has been chosen at the Fermi level (horizontal dashed line). Vertical dotted lines indicate the nuclear positions of the atomic layers of LSMO, and the limits of the CNT.

L is the channel length, v_N is the mean electron velocity in the channel (here, v_F for the CNT), and $\bar{t}_r$ is the mean interfacial transmission coefficient that we estimate later via r_b^*. Equations (1) and (2) hold for ballistic transmission from source to drain, and also for diffusive transport when r_b^* is sufficiently large[20,21], as we have here.

Equation (1) shows that MR is controlled by two factors: trivially γ, and critically τ_n/τ_{sf}. If this ratio were large, the MR would tend to zero, whatever γ. From equation (2), we can express this ratio as:

$$\frac{\tau_n}{\tau_{sf}} = \frac{2L}{v_N \bar{t}_r \tau_{sf}} \quad (3)$$

From equation (1), γ and τ_n/τ_{sf} cannot both be extracted from the MR alone (61% at 5 K, 25 mV), but we necessarily have $\gamma \geq 0.62$ as the denominator cannot be smaller than unity. It is possible that $\gamma = 1$ for half-metallic LSMO, but interfacial imperfections lead to smaller values. The maximum value observed in epitaxial magnetic tunnel junctions[4] with LSMO is 0.95. Here we propose a tentative scenario assuming a reasonable value of $\gamma = 0.8$, which gives $\tau_n/\tau_{sf} \approx 2$.

To estimate τ_{sf}, we obtain τ_n from equation (2), with $L = 2\,\mu m$, $v_N = v_F = 0.8 \times 10^6\,m\,s^{-1}$ (ref. 7) and $\bar{t}_r \approx 0.9 \times 10^{-4}$ estimated from r_b^* using the Landauer equation:

$$r_b^* = \frac{h}{4e^2 \bar{t}_r} \quad (4)$$

where the assumption of two spin-degenerate conduction channels in deriving this equation is realistic[7] even for a multiwall CNT. As r_b^* dominates device resistance R, we take $r_b^* = R/2 \approx 75\,M\Omega$. The above yields $\tau_n \approx 60$ ns, and thus $\tau_{sf} \approx 30$ ns. This value is reasonable given the very weak spin–orbit coupling of carbon, and should also apply to other carbon-based molecules. The corresponding spin diffusion length is $l_{sf} = \sqrt{v_F \tau_{sf} \lambda} \approx 50\,\mu m$, assuming a CNT mean free path[7] of $\lambda \approx 100$ nm.

Equivalent calculations with the best value[4] of $\gamma \approx 0.95$ would reduce τ_{sf} by a factor of 7 and shorten l_{sf} by a factor of 2.7. Alternatively, if hole-doping activates 10 rather than 4 CNT channels, τ_{sf} would increase by a factor of 2.5 and l_{sf} would increase by a factor of 1.6. If both scenarios are active, then clearly the former would outcompete the latter.

Purely metallic structures like magnetic multilayers have the advantage of a large carrier velocity and a large $\bar{t}_r \approx 1$, but τ_{sf} is very short so that a large $\Delta R/R$ can be obtained only when L is short—for example, in current-perpendicular-to-the-plane giant magnetoresistance. The long L in a lateral structure forces $\Delta R/R$ to become small, for example[1], ~5%. When the interfaces are tunnel junctions, in lateral structures suitable for gating, the concomitant reduction of $\bar{t}_r$ leads to an even smaller $\Delta R/R$ (for example[22], $\sim 10^{-4}$).

Semiconductors have the advantage of a long[1] τ_{sf}, but the mean velocity v_N is small. For example, n-type GaAs (carrier density $10^{17}\,cm^{-3}$) has a long low-temperature conduction-band spin lifetime of several nanoseconds, but the mean velocity along a channel axis is $\sim 3 \times 10^4\,m\,s^{-1}$, compared to $10^6\,m\,s^{-1}$ in metals or CNTs. Moreover, semiconductor channels require a small $\bar{t}_r$ for efficient spin injection from metals[23–26]. The MR < 1% of lateral semiconductor structures[2] may be increased to ~40% using a small $L \approx 5$–10 nm in vertical structures[27], but these are unsuitable for gating.

The advantage of CNTs is that they combine the long τ_{sf} of semiconductors with the large[7] v_F of metals. This permits our large MR, despite the long $L = 2\,\mu m$ and the small $\bar{t}_r$. In fact, a small $\bar{t}_r$ is necessary here to limit current at high bias. Working at high bias not only avoids Coulomb blockade and level quantization effects, but is in addition a prerequisite for achieving large output signals.

The bias dependence of the 5 K MR (Fig. 4) is reminiscent of LSMO tunnel junctions[28], but we cannot rule out the possible role of CNT energy bands here. Above the (unresolved) classical zero-bias anomaly, there is a plateau out to $V \approx 110$ mV, and then a steep decrease. The persistence of this plateau to ~110 mV permits the associated output signal (that is, the voltage difference between the parallel and antiparallel configurations for the same current) to increase from $V \times MR = 15$ mV at a bias of 25 mV, up to 65 mV at a bias of 110 mV. This figure of 65 mV falls in a suitable range for applications.

Device MR falls with increasing temperature (Fig. 4), but the field dependence is qualitatively unchanged. Our MR persists to 120 K, which, although well below room temperature, is a significant improvement on previous molecular spintronics devices[9–14]. This loss of performance well below the 365 K Curie temperature of bulk LSMO is probably associated with the well known thermal suppression of spin polarization[3]. A similar fall-off in performance in LSMO tunnel junctions[28] is attributed to a reduced interfacial Curie temperature arising from charge transfer or loss of bulk symmetry. Replacing LSMO with a high-Curie-temperature metal such as Co could solve this problem, but previous results[11–14] were limited by interfacial resistances (<1 MΩ) two orders of magnitude smaller than r_b^*, suggesting the need for tunnel barriers (for example, thin

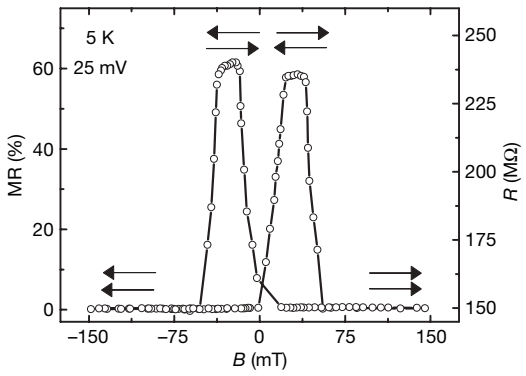


Figure 3 | MR for a LSMO–CNT–LSMO device. Data recorded at 5 K with a bias voltage of 25 mV show two distinct states of resistance R, as the magnetic configuration of the two LSMO electrodes is switched by an applied magnetic field B. The arrows indicate the relative magnetic orientation of the electrodes, which possess different switching fields because of their different widths. The data points and interconnecting lines were generated by averaging over 25 cycles; MR(%) was calculated as $MR(\%) = 100 \times [R(B) - R(0)]/R(B)$. In Supplementary Information, we show similar MR data for three other working devices. One of these three devices was fabricated with silica between the manganite electrodes to prevent the possibility of the CNT sagging. For another of these three devices, data were collected from a single field sweep.

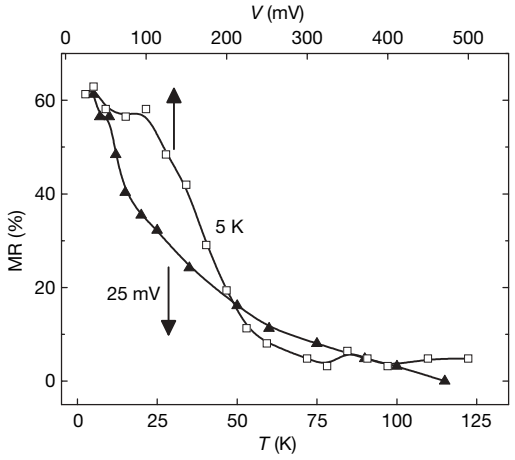


Figure 4 | Temperature and bias dependence of peak MR. The magnitude of the two-state switching seen in Fig. 4 is plotted as a function of bias voltage V at low temperature (open squares), and as a function of temperature T at 25 mV (filled triangles). MR was calculated as $MR(\%) = 100 \times [R_{AP} - R_P]/R_P$.

insulating layers) in order to limit the potentially destructive effect of current at biases sufficient to (1) avoid Coulomb blockade and level quantization effects, and (2) achieve large electrical signals.

Our work forms part of the nascent molecular spintronics approach in which it is possible to manipulate spin-polarized electrons in novel environments. However, the weak spin–orbit coupling in carbon precludes the electrically driven magnetic reversal of spins in a CNT-based spin transistor of the type described in ref. 5. Instead, spin precession induced by the local magnetic field from a ferromagnetic gate, that is, the Hänle effect[1], could be used to flip spins in a CNT. Given that the precession angle induced by a transverse field B during time t is $2\mu_B B\,t/\hbar$, our value of τ_n ($\sim$60 ns) suggests that the application of a modest 10 mT field to a small fraction of the length of a CNT (a few tenths of micrometres) would be sufficient to reverse the spin polarization between injection and detection. Note that here we cannot rule out the possibility that weak components of stray field from our LSMO electrodes reduce the MR values that we present. In future, one might seek non-magnetic channels with intermediate levels of spin–orbit coupling in order to permit spin manipulation by the electric field of a gate without unduly reducing the spin lifetime and the output signals.

METHODS

Experimental. Epitaxial LSMO thin films were grown on closely lattice matched orthorhombic $NdGaO_3$ (001) substrates by pulsed laser deposition with a KrF excimer laser (248 nm, 1 Hz, 2.5 J cm^{-2}, 775 °C, 15 Pa O_2, target–substrate distance = 8 cm). The films display step-terrace growth, and possess in-plane uniaxial magnetocrystalline anisotropy in the orthorhombic [100] direction. Below 360 K the films are ferromagnetic (3.6 μ_B per Mn at 10 K), and on cooling the resistivity decreases to $\sim$60 μΩ cm at 10 K. Using conventional photolithography, electrode tracks (widths 1–4 μm, separation 1.5 μm) were defined perpendicular to [100], so that their magnetizations could be switched independently by an external magnetic field. Multiwall CNTs of diameter $\sim$20 nm grown by arc-discharge (Iljin Nanotech) were subsequently dispersed from a 1,2-dicloroethane solution. A scanning electron microscope was used to confirm the presence of a single nanotube running between adjacent electrically connected electrodes. Electrical measurements of interest were made using a Keithley source meter in constant voltage mode.

Theoretical. First-principles electronic-structure calculations were performed within the density-functional-theory (DFT) framework[19] in the spin-polarized generalized-gradient approximation, using the SIESTA method[29]. Further details on the performance of the method for LSMO can be found elsewhere[30]. The MnO_2-terminated (001) surface of LSMO was described by a 23-layer slab of LSMO, in which one third of the La atoms where replaced[30] by Sr. A (6,6) single-wall CNT was put onto the LSMO surface in a commensurate arrangement in which three unit cells of the CNT were laid along the (100) direction on a 4×2 lateral supercell of LSMO. The mismatch strain is 5%. The atomic positions of the CNT on the previously relaxed surface were obtained by minimizing the mutual DFT forces. Even though experiments were performed on multiwall nanotubes, which are arguably better described in the graphitic limit, we have nevertheless considered a nanotube, as the dimensionality greatly affects the contact resistance, and the qualitative picture emerging from the calculations should remain.

Received 7 September; accepted 27 November 2006.

1. Žutić, I., Fabian, J. & das Sarma, S. Spintronics: Fundamentals and applications. *Rev. Mod. Phys.* **76,** 323–410 (2004).
2. Jonker, B. T. & Flatté, M. E. F. in *Nanomagnetism* (eds Mills, D. L. & Bland, J. A. C.) 227–272 (Elsevier, Amsterdam, 2006).
3. Park, J.-H. *et al.* Direct evidence for a half-metallic ferromagnet. *Nature* **392,** 794–796 (1998).
4. Bowen, M. *et al.* Nearly total spin-polarization in $La_{2/3}Sr_{1/3}MnO_3$ from tunnelling experiments. *Appl. Phys. Lett.* **82,** 233–235 (2003).
5. Datta, S. & Das, B. Electric analog of the electro-optic modulator. *Appl. Phys. Lett.* **56,** 665–667 (1990).
6. Dresselhaus, M. S., Dresselhaus, G. & Avouris, Ph (eds) *Carbon Nanotubes* (Springer, Berlin, 2001).
7. Buitelaar, M. R., Bachtold, A., Nussbaumer, T., Iqbal, M. & Schönenberger, C. Multiwall carbon nanotubes as quantum dots. *Phys. Rev. Lett.* **88,** 156801 (2002).
8. Joachim, C., Gimzewski, J. K. & Aviram, A. Electronics using hybrid-molecular and mono-molecular devices. *Nature* **408,** 541–548 (2000).
9. Petta, J. R., Slater, S. K. & Ralph, D. C. Spin-dependent transport in molecular tunnel junctions. *Phys. Rev. Lett.* **93,** 136601 (2004).
10. Pasupathy, A. N. *et al.* The Kondo effect in the presence of ferromagnetism. *Science* **306,** 86–89 (2004).
11. Tsukagoshi, K., Alphenaar, B. W. & Ago, H. Coherent transport of electron spin in a ferromagnetically contacted carbon nanotube. *Nature* **401,** 572–574 (1999).
12. Sahoo, S. *et al.* Electric field control of spin transport. *Nature Phys.* **1,** 99–102 (2005).
13. Jensen, A., Hauptmann, J. R., Nygård, J. & Lindelof, P. E. Magnetoresistance in ferromagnetically contacted single-wall carbon nanotubes. *Phys. Rev. B* **72,** 035419 (2005).
14. Tombros, N., van der Molen, S. J. & van Wees, B. J. Separating spin and charge transport in single-wall carbon nanotubes. *Phys. Rev. B* **73,** 233403 (2006).
15. Meservey, R. & Tedrow, P. M. Spin-polarized electron tunneling. *Phys. Rep.* **238,** 173–243 (1994).
16. Jorgensen, H. I., Grove-Rasmussen, K., Novotny, T., Flensberg, K. & Lindelof, P. E. Electron transport in single-wall carbon nanotube weak links in the Fabry-Perot regime. *Phys. Rev. Lett.* **96,** 207003 (2006).
17. Mieville, L., Wordledge, D., Geballe, T. H., Contreras, R. & Char, K. Transport across conducting ferromagnetic oxides/metal interfaces. *Appl. Phys. Lett.* **73,** 1736–1739 (1998).
18. Hueso, L. E. *et al.* Electrical transport between epitaxial manganites and carbon nanotubes. *Appl. Phys. Lett.* **88,** 083120 (2006).
19. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlations effects. *Phys. Rev.* **140,** 1133–1138 (1965).
20. George, J. M. *et al.* Electrical spin injection in GaMnAs-based junctions. *Mol. Phys. Rep.* **40,** 23–33 (2004).
21. Fert, A., George, J. M., Jaffrès, H. & Mattana, R. Semiconductors between spin-polarized source and drain. *IEEE Trans. Electron. Devices* (special issue on spintronics) (in the press); preprint at ⟨http://arxiv.org/abs/cond-mat/0612495⟩ (2006).
22. Jedema, F. J., Heersche, H. B., Filip, A. T., Baselmans, J. J. A. & van Wees, B. J. Electrical detection of spin precession in a metallic mesoscopic spin valve. *Nature* **416,** 713–716 (2002).
23. Schmidt, G., Ferrand, D., Molenkamp, L. W., Filip, A. T. & van Wees, B. J. Fundamental obstacle for electrical spin injection from a ferromagnetic metal into a diffusive semiconductor. *Phys. Rev. B* **62,** 4790–4793 (2000).
24. Rashba, E. Theory of electrical spin injection: tunnel contacts as a solution of the conductivity mismatch problem. *Phys. Rev. B* **62,** 16267–16270 (2000).
25. Fert, A. & Jaffrès, H. Conditions for efficient spin injection from a ferromagnetic metal into a semiconductor. *Phys. Rev. B* **64,** 184420 (2001).
26. Smith, D. L. & Silver, R. N. Electrical spin injection into semiconductors. *Phys. Rev. B* **64,** 045323 (2001).
27. Mattana, R. *et al.* Electrical detection of spin accumulation in a *p*-type GaAs quantum well. *Phys. Rev. Lett.* **90,** 166601 (2003).
28. Bowen, M. *et al.* Spin-polarized tunnelling spectroscopy in tunnel junctions with half-metallic electrodes. *Phys. Rev. Lett.* **95,** 137203 (2005).
29. Soler, J. M. *et al.* The SIESTA method for ab initio order-N materials simulation. *J. Phys. Condens. Matter* **14,** 2745–2779 (2002).
30. Ferrari, V., Pruneda, J. M. A. & Artacho, E. Density functionals and half-metallicity in $La_{2/3}Sr_{1/3}MnO_3$. *Phys. Stat. Sol. a* **203,** 1437–1441 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank G. A. J. Amaratunga, M. Bibes, H. Bouchiat, L. Brey, M. R. Buitelaar, M. J. Calderón, S. N. Cha, M. Chhowalla, A. Cottet, H. Jaffrès, D.-J. Kang, T. Kontos, P. Seneor and N. A. Spaldin. This work was funded by the UK EPSRC, NERC, BNFL, The Royal Society, the Spanish MEC (J.M.P.), Donostia International Physics Center (E.A.) and the EU.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to N.D.M. (ndm12@cam.ac.uk).

LETTERS

A 160-kilobit molecular electronic memory patterned at 10^{11} bits per square centimetre

Jonathan E. Green[1]*, Jang Wook Choi[1]*, Akram Boukai[1], Yuri Bunimovich[1], Ezekiel Johnston-Halperin[1]†, Erica DeIonno[1], Yi Luo[1]†, Bonnie A. Sheriff[1], Ke Xu[1], Young Shik Shin[1], Hsian-Rong Tseng[2]†, J. Fraser Stoddart[2] & James R. Heath[1]

The primary metric for gauging progress in the various semiconductor integrated circuit technologies is the spacing, or pitch, between the most closely spaced wires within a dynamic random access memory (DRAM) circuit[1]. Modern DRAM circuits have 140 nm pitch wires and a memory cell size of 0.0408 μm^2. Improving integrated circuit technology will require that these dimensions decrease over time. However, at present a large fraction of the patterning and materials requirements that we expect to need for the construction of new integrated circuit technologies in 2013 have 'no known solution'[1]. Promising ingredients for advances in integrated circuit technology are nanowires[2], molecular electronics[3] and defect-tolerant architectures[4], as demonstrated by reports of single devices[5–7] and small circuits[8,9]. Methods of extending these approaches to large-scale, high-density circuitry are largely undeveloped. Here we describe a 160,000-bit molecular electronic memory circuit, fabricated at a density of 10^{11} bits cm^{-2} (pitch 33 nm; memory cell size 0.0011 μm^2), that is, roughly analogous to the dimensions of a DRAM circuit[1] projected to be available by 2020. A monolayer of bistable, [2]rotaxane molecules[10] served as the data storage elements. Although the circuit has large numbers of defects, those defects could be readily identified through electronic testing and isolated using software coding. The working bits were then configured to form a fully functional random access memory circuit for storing and retrieving information.

The 'crossbar' geometry—a periodic array of crossed wires—provides a promising architecture for nanoelectronic circuitry[11–15], as was experimentally demonstrated by the Teramac supercomputer[4]. The crossbar is tolerant of manufacturing defects—a trait that becomes increasingly important as devices approach macromolecular dimensions and non-traditional (and imperfect) fabrication methods are employed. For example, Teramac had nearly a quarter of a million hardware defects and yet could be configured into a robust computing machine. The crossbar geometry is similar in structure to a two-dimensional crystal, implying that non-traditional methods can be employed for its construction[9,16,17]. The crossbar is also the highest-density two-dimensional digital circuit for which every device can be independently addressed[4]. This attribute enables the circuit to be fully tested for manufacturing defects and to be subsequently configured into a working circuit.

A few groups have reported on non-lithographic methods for fabricating crossbar circuits[16,18], but most methods are not yet feasible for fabricating more than a handful of devices. Furthermore, the assembly of nanowires into narrow-pitch crossbars without electrically shorting adjacent nanowires remains a challenge. We previously reported on the superlattice nanowire pattern transfer (SNAP) method for producing ultradense, highly aligned arrays and crossbars of high-aspect-ratio metal or semiconductor nanowires[19] containing up to 1,400 nanowires at a pitch as small as 15 nm (see Supplementary Information). We also reported on the use of bistable [2]rotaxane molecular monolayers as the storage elements within crossbar memories, using micrometre-scale wiring[20]. Here we combine these methods and materials, along with the defect-tolerance concepts learned from the Teramac supercomputer, to construct and test a memory circuit at extreme dimensions: the entire 160,000-bit crossbar is approximately the size of a white blood cell ($\sim 13 \times 13\ \mu m^2$).

The fabrication of this molecular memory circuit, which required the integration of molecular switches with large numbers of semiconductor and metal nanowires, presented a number of challenges. We needed to develop a process flow in which the [2]rotaxane molecular monolayer was incorporated into the circuit as close to the final step as possible, and then protect that monolayer during subsequent processing steps. We also had to establish electronic measurement protocols that could be used to follow the conductivity status of the nanowires during the entire nanofabrication procedure. Details of this process flow, along with the various electronic testing protocols, are presented in the Supplementary Information.

The assembled crossbar memory (Fig. 1) consisted of 400 Si bottom-nanowire electrodes (16 nm wide, 33 nm pitch; phosphorus-doped, $n = 5 \times 10^{19}$ cm^{-3}) crossed by 400 Ti top-nanowire electrodes (16 nm wide, 33 nm pitch), sandwiching a monolayer of bistable [2]rotaxanes (Fig. 2). Each bit corresponds to an individual molecular switch tunnel junction (MSTJ) defined by a Si bottom nanowire and Ti top nanowire and contained approximately 100 [2]rotaxane molecules. Electrical contacts were established to several bottom and top nanowires to allow us to test up to 180 effective bits ('ebits') from the central region of the crossbar, but only 128 were actually tested, owing to measurement constraints. Because SNAP nanowires are patterned beyond the resolution of lithographic methods[21], each test electrode contacted two to four nanowires. (Fig. 1b). We recently reported on a demultiplexer that would allow for this memory circuit to be fully tested[22]; however, implementation of that demultiplexer would have added significant complexity to an already demanding procedure, and wasn't necessary to demonstrate the viability of this circuit. The 128 tested ebits represented between 0.5–0.7% of the full 160-kilobit crossbar distributed over 6% of the circuit area

[1]Division of Chemistry and Chemical Engineering and the Kavli Nanoscience Institute, Caltech, Pasadena, California 91125, USA. [2]California NanoSystems Institute and the Department of Chemistry and Biochemistry, University of California at Los Angeles, 405 Hilgard Avenue, Los Angeles, California 90095-1569, USA. †Present addresses: Department of Electrical and Computer Engineering, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213, USA (Y.L.); Crump Institute for Molecular Imaging, University of California, Los Angeles, California 90095, USA (H.-R.T.); Department of Physics, Ohio State University, 191 W. Woodruff Ave. Columbus, OH 43210-1117 (E. J.-H.)
*These authors contributed equally to this work.

(Fig. 1c). We believe that this relatively small portion of the crossbar is representative of the overall circuit, on the basis of results from testing four other similarly prepared circuits.

By scanning electron microscopy (SEM) inspection, the crossbar appeared to be structurally defect-free, with no evidence of broken, wandering or electrically shorted nanowires. Nevertheless, electrical testing identified a large number of defective bits and the nature of those defects. This testing was done by first applying a +1.5 V pulse relative to the Si bottom-nanowire electrodes for 0.2 s to set all bits to '1', and then reading each ebit sequentially using a non-perturbing +0.2 V bias. A −1.5 V, 0.2 s pulse was then applied to set all bits to '0'. The status of each of the ebits was again read. The 1/0 current ratios are presented in Fig. 3a. About 50% of the bits yielded some sort of switching response. Some of that response, however, may have originated from parasitic current pathways through the crossbar array. This is an inherent drawback of crossbar architectures wherein each junction is electrically connected to every other junction. The standard remedy is to incorporate diodes at each crosspoint[23], and although the molecule/Ti interface yields some rectification[24], we additionally grounded all nanowire electrodes not being used during a read or write step. We established a threshold for a 'good' bit based upon a minimum 1/0 current ratio of ~1.5. About 25% of the ebits passed this threshold. Electrical testing revealed several types of defects (Fig. 3b). The defects classified as 'switch defects' probably

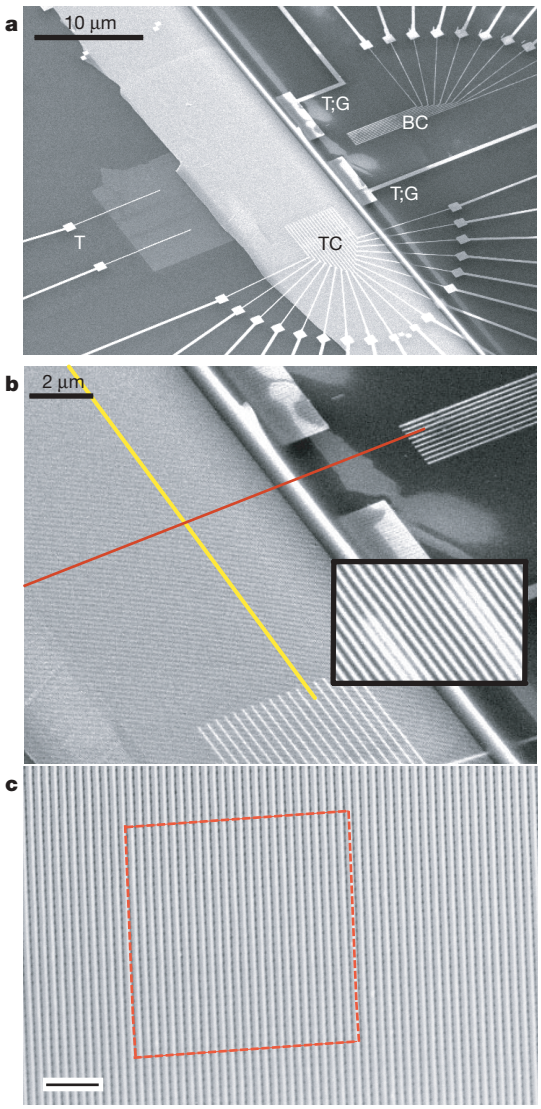


Figure 1 | SEMs of the nanowire crossbar memory. **a**, Image of the entire circuit. The array of 400 Si bottom nanowires is seen as the light grey rectangular patch extending diagonally up from bottom left. The top array of 400 Ti nanowires is covered by the SNAP template of 400 Pt nanowires, and extends diagonally down from top left. Testing contacts (T) are for monitoring the electrical properties of the Si nanowires during the fabrication steps. Two of those contacts are also grounding contacts (G), and are used for grounding most of the Si nanowires during the memory evaluation, writing and reading steps. Eighteen electron-beam-lithography patterned top contacts (TC) and ten such bottom contacts (BC) are also visible. The scale bar is 10 μm. **b**, An SEM image showing the cross-point of top- (red) and bottom- (yellow) nanowire electrodes. Each cross-point corresponds to an ebit in memory testing. The electron-beam-lithography defined contacts bridged two to four nanowires each (inset). The scale bar is 2 μm. **c**, High-resolution SEM of approximately 2,500 junctions out of a 160,000-junction nanowire crossbar circuit. The red square highlights an area of the memory that is equivalent to the number of bits that were tested. The scale bar is 200 nm.

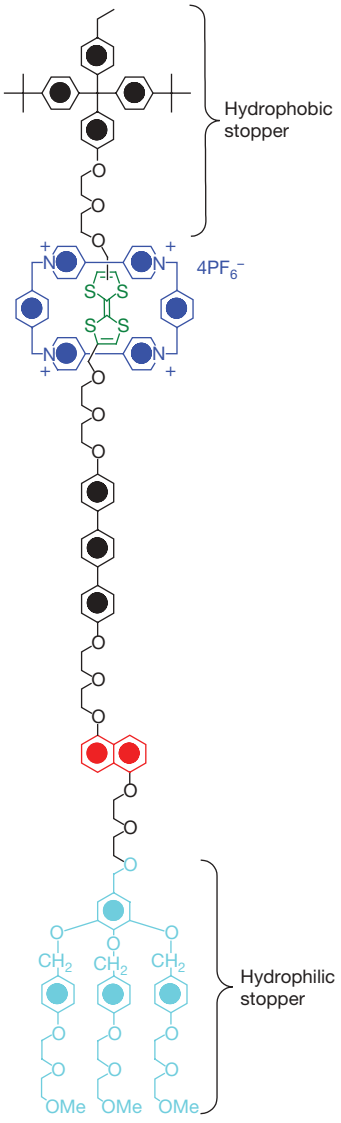


Figure 2 | Structural formula of the bistable [2]rotaxane used in the crossbar memory. The ground-state conformation is shown and corresponds to the low-conductance, or '0' co-conformation. The molecule is oriented with the (light blue) hydrophilic stopper in contact with the Si bottom-nanowire electrodes. The switching mechanism involves oxidation of the (green) tetrathiafulvalene (TTF) site to the TTF^{+1} or TTF^{+2} oxidation state, followed by translation of the blue ring from the TTF^{+} site to the (red) dioxynapthalene site. The TTF^{+} is reduced back to the TTF^{0} oxidation state to form the metastable state co-conformer, which is the high-conductance, or '1' state. The metastable state will relax back to the ground state with a half-life of about an hour.

arose from subnanometre variations in the reactive ion etching process that was used to define the Ti crossbar top nanowires. Isolated devices, or crossbar memories patterned at substantially lower densities and with larger wires, can typically be prepared with a nearly 100% yield. The switch defects led to only a proportional loss in the yield of functional bits, whereas bad contacts or shorted nanowires removed an entire row of bits from operation.

An important result from the defect map (Fig. 3b) is that the good and bad bits are randomly dispersed, implying that the crossbar junctions are operationally independent of one another. However, the ultimate test of any memory is whether it can be used to store and retrieve information. From the defect map, we identified the addresses of the usable ebits, and from those addresses configured an operational memory (Fig. 4a).

The solid-state switching signature of the bistable [2]rotaxanes that were used here has been shown to originate from electrochemically addressable, molecular mechanical switching for certain device structures[10,20], but not for metal wire/molecule/metal wire junctions[25]. In fact, our desire to use molecular mechanical bistable switches as the storage elements is what dictated our choice of the Si nanowire/molecule/Ti nanowire crossbar structure. This switching signature should be effectively size-invariant, meaning that it should scale to the macromolecular dimensions of these crossbar junctions. Solid-state-based switching materials[26,27] will probably not exhibit similar scaling since they arise from inherently bulk properties. The thermodynamic and kinetic parameters describing both the bistability and switching mechanism of the [2]rotaxane switch (and similar molecular mechanical switches[28]) have been quantified in a variety of environments[10]. Those measurements required robust switching devices that could be cycled many times and at various temperatures. The memory bits measured here were much more delicate—although all good ebits could be cycled multiple times (as shown by the testing and writing steps), most ebits failed after a half-dozen or so cycles, and none lasted longer than ten cycles. However, we measured the rate of relaxation from the 1→0 state for many of the ebits (Fig. 4b). From a device perspective, this represents the volatility, or memory retention time, of the bits. With respect to the bistable [2]rotaxane switching cycle, this represents a measurement of the rate-limiting kinetic step within the switching cycle[10]. Our measured rate (90 ± 40 min; median decay 75 min) was statistically equivalent to that reported for much larger (and more fully characterized)

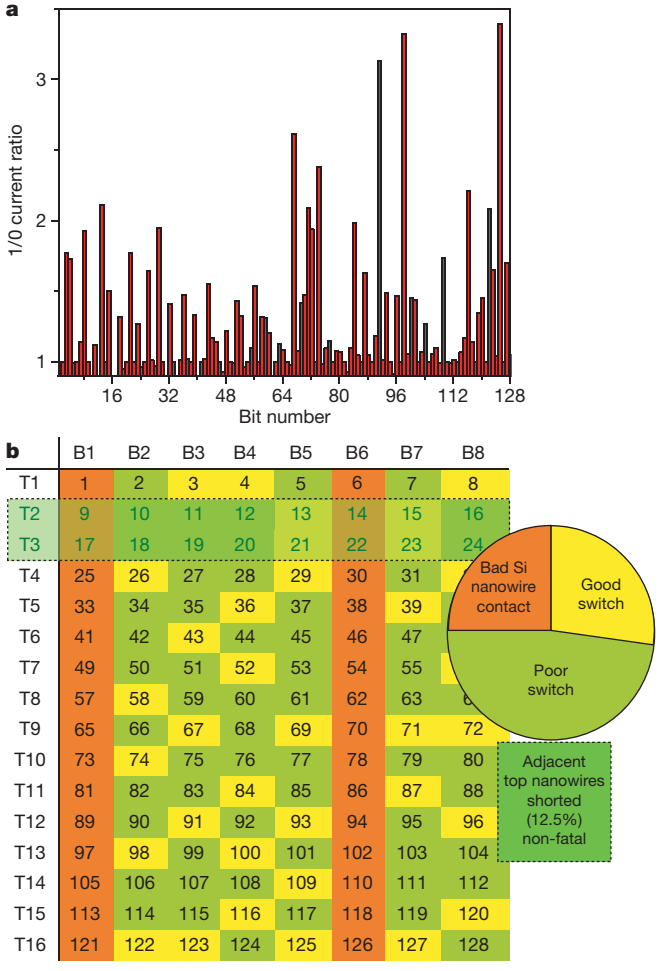


Figure 3 | Data from evaluating the performance of the 128 ebits within the crossbar memory circuit. a, The current ratio of the 1 state divided by the 0 state of the tested ebits. Note that many of the ebits exhibit little to no switching response. Those ebits are defective. **b**, A map of the defective and useable ebits, along with a pie-chart giving the testing statistics. Note that, except for the bad Si nanowire contacts on bottom electrodes B1 and B6, and the shorted top electrodes T2 and T3, the defective and good bits are randomly distributed. Poor switch defects are poorly switching or non-switching ebits that either exhibited an open-circuit conductance (26% of all ebits tested) or a conductance similar to that of a closed bit (22% of all ebits tested).

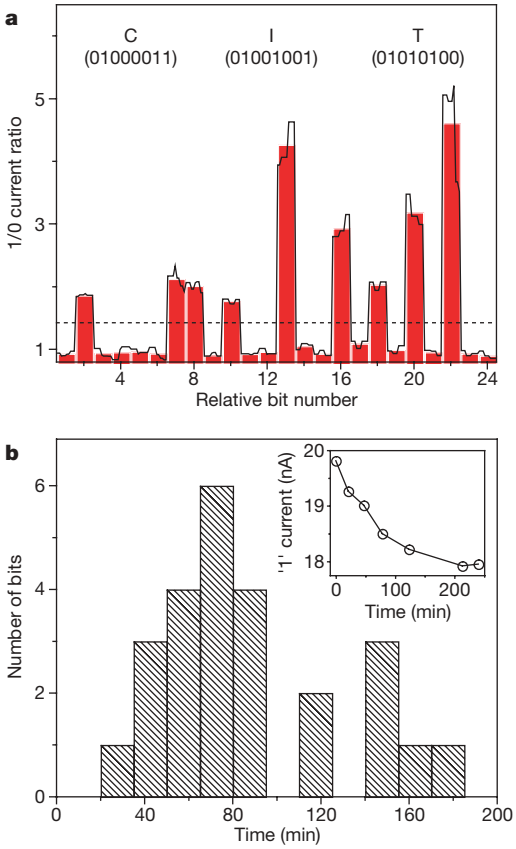


Figure 4 | Demonstration of memory storage and retention characteristics from the molecular electronic crossbar memory. a, A demonstration of point-addressability within the crossbar. Good ebits were selected from the defect mapping of the tested portion of the crossbar. A string of 0s and 1s corresponding to the ASCII characters for 'CIT' (the abbreviation for the California Institute of Technology) were stored and read out sequentially. The dotted line indicates the separation between the 0 and 1 states of the individual ebits. The black trace is raw data showing ten sequential readings of each bit, while the red bars represent the average of those ten readings. Note that deviations of individual readings from their average are well separated from the threshold 1/0 line. **b**, A histogram representing the 1/e decay time of the 1 state to the 0 state. The 25 ebits represented in the data were each large ebits, comprising approximately 100 junctions, to increase the measurement signal-to-noise ratio. Raw data from a single large ebit is shown in the inset. The solid line is a guide to the eye, not a fit.

devices (58 ± 5 min)[10]. Thus, our results are consistent with a molecular mechanism for the switching operation[10,20].

Many scientific and engineering challenges, such as device robustness, improved etching tools and improved switching speed, remain to be addressed before the type of crossbar memory described here can be practical. Nevertheless, this 160,000-bit molecular memory does indicate that at least some of the most challenging scientific issues associated with integrating nanowires, molecular materials, and defect-tolerant circuit architectures at extreme dimensions are solvable. Although it is unlikely that these digital circuits will scale to a density that is only limited by the size of the molecular switches, it should be possible to increase the bit density considerably over what is described here. Recent nano-imprinting results suggest that high-throughput manufacturing of these types of circuits may be possible[29]. Finally, these results provide a compelling demonstration of many of the nanotechnology concepts that were introduced by the Teramac supercomputer several years ago, albeit using a circuit that contained a significantly higher fraction of defective components than did the Teramac machine[4].

Received 18 July; accepted 16 November 2006.

1. *The International Technology Roadmap for Semiconductors (ITRS)*: process integration, devices, and structures. (Semiconductor Industry Association, San Jose, California, 2005). ⟨http://www.itrs.net/reports.html⟩.
2. Yang, P. & Kim, F. Langmuir-Blodgett assembly of one-dimensional nanostructures. *ChemPhysChem* **3,** 503–506 (2002).
3. Heath, J. R. & Ratner, M. A. Molecular electronics. *Phys. Today* **56,** 43–49 (2003).
4. Heath, J. R., Kuekes, P. J., Snider, G. S. & Williams, R. S. A defect-tolerant computer architecture: Opportunities for nanotechnology. *Science* **280,** 1716–1721 (1998).
5. Xiang, J. *et al.* Ge/Si nanowire heterostructures as high-performance field-effect transistors. *Nature* **441,** 489–493 (2006).
6. Zhou, X., Park, J. Y., Huang, S., Liu, J. & McEuen, P. L. Band structure, phonon scattering, and the performance limit of single-walled carbon nanotube transistors. *Phys. Rev. Lett.* **95,** 146805 (2005).
7. Pop, E. *et al.* Negative differential conductance and hot phonons in suspended nanotube molecular wires. *Phys. Rev. Lett.* **95,** 155505 (2005).
8. Chen, Z. *et al.* An integrated logic circuit assembled on a single carbon nanotube. *Science* **311,** 1735 (2006).
9. Zhong, Z. H., Wang, D. L., Cui, Y., Bockrath, M. W. & Lieber, C. M. Nanowire crossbar arrays as address decoders for integrated nanosystems. *Science* **302,** 1377–1379 (2003).
10. Choi, J. W. *et al.* Ground-state equilibrium thermodynamics and switching kinetics of bistable [2]rotaxanes switched in solution, polymer gels, and molecular electronic devices. *Chem. Eur. J.* **12,** 261–279 (2006).
11. DeHon, A. & Naeimi, H. Seven strategies for tolerating highly defective fabrication. *IEEE Design Test Comput.* **22,** 306–315 (2005).
12. Lee, M. H., Kim, Y. K. & Choi, Y. H. A defect-tolerant memory architecture for molecular electronics. *IEEE Trans. Nanotechnol.* **3,** 152–157 (2004).
13. DeHon, A., Goldstein, S. C., Kuekes, P. J. & Lincoln, P. Nonphotolithographic nanoscale memory density prospects. *IEEE Trans. Nanotechnol.* **4,** 215–228 (2005).
14. Snider, G., Kuekes, P., Hogg, T. & Williams, R. S. Nanoelectronic architectures. *Appl. Phys. A* **80,** 1183–1195 (2005).
15. Stan, M. R., Franzon, P. D., Goldstein, S. C., Lach, J. C. & Ziegler, M. M. Molecular electronics: From devices and interconnect to circuits and architecture. *Proc. IEEE* **91,** 1940–1957 (2003).
16. Diehl, M., Beckman, R., Yaliraki, S. & Heath, J. R. Self-assembly of deterministic carbon nanotube wiring networks. *Angew. Chem. Int. Edn Engl.* **41,** 353–356 (2002).
17. Wu, W. *et al.* One-kilobit cross-bar molecular memory circuits at 30-nm half-pitch fabricated by nanoimprint lithography. *Appl. Phys. A* **80,** 1173–1178 (2005).
18. Huang, Y., Duan, X. F., Wei, Q. Q. & Lieber, C. M. Directed assembly of one-dimensional nanostructures into functional networks. *Science* **291,** 630–633 (2001).
19. Melosh, N. A. *et al.* Ultrahigh-density nanowire lattices and circuits. *Science* **300,** 112–115 (2003).
20. Luo, Y. *et al.* Two-dimensional molecular electronics circuits. *ChemPhysChem* **3,** 519–525 (2002).
21. Vieu, C. *et al.* Electron beam lithography: resolution limits and applications. *Appl. Surf. Sci.* **164,** 111–117 (2000).
22. Beckman, R., Johnston-Halperin, E., Luo, Y., Green, J. E. & Heath, J. R. Bridging dimensions: demultiplexing ultrahigh-density nanowire circuits. *Science* **310,** 465–468 (2005).
23. Parkin, S. S. P. *et al.* Exchange-biased magnetic tunnel junctions and application to nonvolatile magnetic random access memory. *J. Appl. Phys.* **85,** 5828–5833 (1999).
24. McCreery, R. L. Molecular electronic junctions. *Chem. Mater.* **16,** 4477–4496 (2004).
25. Chen, Y. *et al.* Nanoscale molecular-switch crossbar circuits *Nanotechnology* **14,** 462–468 (2003).
26. Allwood, D. A. *et al.* Magnetic domain-wall logic. *Science* **309,** 1688–1692 (2005).
27. Waser, R. & Rudiger, A. Ferroelectrics—pushing towards the digital storage limit. *Nature Mater.* **3,** 81–82 (2004).
28. Katz, E., Baron, R., Willner, I., Richke, N. & Levine, R. D. Temperature-dependent and friction-controlled electrochemically induced shuttling along molecular strings associated with electrodes. *ChemPhysChem* **6,** 2179–2189 (2005).
29. Jung, G. Y. *et al.* Circuit fabrication at 17 nm half-pitch by nanoimprint lithography. *Nano Lett.* **6,** 351–354 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported primarily by the DARPA MolApps Program with additional support from the MARCO Center for Advanced Materials and Devices and the National Science Foundation. J.W.C. and Y.S.S. acknowledge fellowships from the Samsung Corporation. We are grateful to Y. Liu and S. Saha for preparing the [2]rotaxane molecule used in this work.

Author Contributions The [2]rotaxane molecular switches were designed and originally synthesized by H.-R.T. and J.F.S. All other authors contributed to the design, fabrication and testing of the memory circuit.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.R.H. (heath@caltech.edu).

LETTERS

Inconsistent correlation of seismic layer 2a and lava layer thickness in oceanic crust

Gail L. Christeson[1], Kirk D. McIntosh[1] & Jeffrey A. Karson[2]

At mid-ocean ridges with fast to intermediate spreading rates, the upper section of oceanic crust is composed of lavas overlying a sheeted dyke complex. These units are formed by dykes intruding into rocks overlying a magma chamber, with lavas erupting at the ocean floor. Seismic reflection data acquired over young oceanic crust commonly image a reflector known as 'layer 2A', which is typically interpreted as defining the geologic boundary between lavas and dykes[1–3]. An alternative hypothesis is that the reflector is associated with an alteration boundary within the lava unit[4–6]. Many studies have used mapped variability in layer 2A thickness to make inferences regarding the geology of the oceanic crust, including volcanic construction, dyke intrusion and faulting[7–10]. However, there has been no link between the geologic and seismological structure of oceanic crust except at a few deep drill holes. Here we show that, although the layer 2A reflector is imaged near the top of the sheeted dyke complex at fast-spreading crust located adjacent to the Hess Deep rift, it is imaged significantly above the sheeted dykes section at intermediate-spreading crust located near the Blanco transform fault. Although the lavas and underlying transition zone thicknesses differ by about a factor of two, the shallow seismic structure is remarkably similar at the two locations. This implies that seismic layer 2A cannot be used reliably to map the boundary between lavas and dykes in young oceanic crust. Instead we argue that the seismic layer 2A reflector corresponds to an alteration boundary that can be located either within the lava section or near the top of the sheeted dyke complex of oceanic crust.

Fault scarps along the north wall at the Hess Deep rift (HDR) and the Blanco transform fault (BTF) expose *in situ* cross-sections of the oceanic crust parallel to the spreading direction over lateral distances of tens of kilometres and vertical distances of up to 2 km. The HDR (Fig. 1a) is a broad rift valley near the tip of the westward-propagating Cocos–Nazca plate boundary in the equatorial Pacific Ocean[11,12]. Along the north wall of the HDR, the uppermost 2 km of 1-Myr-old crust generated at the fast-spreading ($\sim$65 mm yr^{-1} half-spreading rate) East Pacific Rise is exposed and has been extensively mapped using manned submersibles and remotely operated vehicles[13–15]. The 360-km-long BTF links the Gorda ridge to the south and the Juan de Fuca ridge to the north (Fig. 1b). The northern wall of the western segment of this transform is stair-stepped with major cliff faces, providing continuous outcrops up to hundreds of metres high. The exposed crust was generated near the southern end of the Juan de Fuca ridge at an intermediate spreading rate of $\sim$30 mm yr^{-1} (half-rate)[16], and has been studied with a series of manned submersible dives[17–19].

Generalized cross-sections for the mapped sections of the HDR and BTF scarps are illustrated in Fig. 2. The lava unit at both locations contains essentially intact pillow lavas with lesser lobate flows and rare tabular sheet flows; the upper lavas are subhorizontal and undeformed or weakly fractured. The deeper lava flows dip towards the spreading centre, and are more fractured and altered[15,19]. The lava unit is underlain by a transition zone containing lavas and dykes; fracturing and alteration are heterogeneous but pervasive in this layer[15,19]. The lava unit and transition zone are underlain by a sheeted dyke complex where the dykes predominantly dip away from the spreading centre where they were formed[15,19]. All units are variable in thickness, but thicknesses are greater at the BTF than at the HDR (Fig. 2). The average thicknesses of the lava units exposed at the HDR and the BTF are 300 m and 450 m, respectively; the average thicknesses of the transition zone are 150 m and 700 m, respectively. The mean thickness of the lava unit and transition zone together is more than twice as much at the intermediate-spreading crust exposed at the BTF (1,150 m) compared to the fast-spreading crust exposed at the HDR scarp (450 m).

The seismic structure of oceanic crust is divided into layer 2 and layer 3, and layer 2 is typically subdivided further into layers 2A and 2B. Seismic layer 2A consists of a low-seismic-velocity ($<$3.0 km s^{-1}) layer directly beneath the sea floor, underlain by a high-gradient region (where seismic velocities $>$4–5 km s^{-1} are reached in a few hundred metres[5,6,20]); seismic layer 2B is the underlying layer, with velocities $>$4–5 km s^{-1}. Wide-angle reflections from the high-gradient region at the base of layer 2A, when processed correctly in multichannel seismic data[3,21], can be used to image the layer 2A/2B boundary in oceanic crust. We conducted nearly identical seismic experiments adjacent to the north walls of both the HDR and BTF scarps (Fig. 1) for the purpose of directly correlating the seismic layer 2A/2B boundary with mapped geologic units in young oceanic crust. At the HDR, 16 scarp-parallel and 7 scarp-perpendicular multichannel seismic reflection profiles were acquired, with spacing between the scarp-parallel profiles of 0.5–2.0 km, and spacing between the scarp-perpendicular profiles averaging 7.5 km. A similar survey was conducted near the BTF, with 10 scarp-parallel profiles at 0.5–2.0 km spacing and 4 scarp-perpendicular profiles at 7.5 km spacing. Data were acquired on board the R/V *Maurice Ewing*, using a 6-km-long hydrophone cable with 480 channels; the seismic source was an array of 10 air guns with a total volume of 3,050 cubic inches. Processing for imaging the layer 2A event followed closely the method described in ref. 3, except that we also applied a parabolic 'Radon' transform to the sorted data for removal of multiple energy. Sample sections are displayed in Fig. 3.

A layer 2A event is imaged on all profiles, although it is typically intermittent in nature, especially near the scarps (Fig. 3). The two-way travel time (TWTT) between the sea floor and the layer 2A event varies from 0.1 to 0.8 s, with lateral variability in TWTT to the layer 2A event occurring over a shorter wavelength in the scarp-parallel direction (Fig. 3a and c) compared to the scarp-perpendicular

[1]University of Texas Institute for Geophysics, Jackson School of Geosciences, J.J. Pickle Research Campus, Mail Code R2200, 10100 Burnet Road, Austin, Texas 78758, USA.
[2]Department of Earth Sciences, Syracuse University, Syracuse, New York 13244, USA.

direction (Fig. 3b and d). The mean TWTT between the sea floor and the layer 2A event is similar between the two survey regions: 0.37 s at the HDR and 0.36 s at the BTF. These values are comparable to the 0.3–0.45 s TWTT to the layer 2A event observed on the flanks of fast-spreading and intermediate-spreading ridges[3,22]. We convert the TWTT between the sea floor and the layer 2A event to layer 2A thickness using interval velocities of 2,600 m s^{-1} at the HDR and 2,700 m s^{-1} at the BTF; details of our velocity analyses can be found in Supplementary Information. Our estimated mean layer 2A thicknesses of 0.48 km at the HDR and 0.49 km at the BTF are remarkably similar, despite the mapped differences in geologic structure (Table 1).

As expected, the imaging environment is poor at the scarp edge and we cannot image the layer 2A event at the exact locations of the submersible observations of the geologic structure. However, our grid of seismic profiles indicates that lateral variability in TWTT between the sea floor and the layer 2A event is low in the scarp-perpendicular direction. This suggests that the layer 2A event imaged on the profiles near the scarp edge is representative of the structure at the submersible observation locations. In Fig. 2, travel times to the layer 2A event for the five scarp-parallel profiles within 2.5 km of the scarp edge have been converted to depth and projected onto the geologic cross-sections. At the HDR, the seismic layer 2A/2B boundary is generally located within, but near the top of, the sheeted dyke complex. In the eastern half of the BTF cross-section, the seismic layer 2A/2B boundary is located significantly above the sheeted dyke complex, near the boundary between the lava section and the transition zone. In the western half of the BTF cross-section, talus covers the sheeted dyke contact, but its depth is estimated at 3.4–3.5 km from magnetic data[18]; if this depth is correct (which would be consistent with other areas along the scarp), then the layer 2A/2B boundary is located within the transition zone of lavas and dykes in this region as well.

There are two prevalent hypotheses for the nature of the seismic layer 2A/2B boundary; the underlying assumption of both hypotheses is that a decrease in porosity is responsible for the observed velocity increase across this boundary. The first hypothesis is that the layer 2A/2B boundary corresponds to the geologic boundary between lavas and dykes[1–3]. Although both lavas and dykes are composed of basaltic material, the lavas will have a higher porosity, and hence lower seismic velocity, than the dykes, owing to a greater fracture density and volume of void spaces. At the HDR, the layer 2A/2B boundary is located near the top of the sheeted dykes, possibly supporting this hypothesis (Fig. 2a); however, at the BTF the layer 2A/2B boundary is located significantly above (>500 m) the boundary between the lavas and dykes (Fig. 2b). The alternative hypothesis for the nature of the layer 2A/2B boundary is that it corresponds to an alteration boundary within the upper crust, probably in the lava unit[4–6]. Results at Deep Sea Drilling Program (DSDP) Site 504B[23] and Ocean Drilling Program (ODP) Site 1256D[24] show a general increase in temperature and intensity of alteration with depth; hydrothermal mineralization associated with alteration fills cracks, decreases porosities, and increases seismic velocities[4–6]. At the BTF, the seismic layer 2A/2B boundary is generally located within the transition zone (Fig. 2b) and therefore might correspond to an alteration boundary

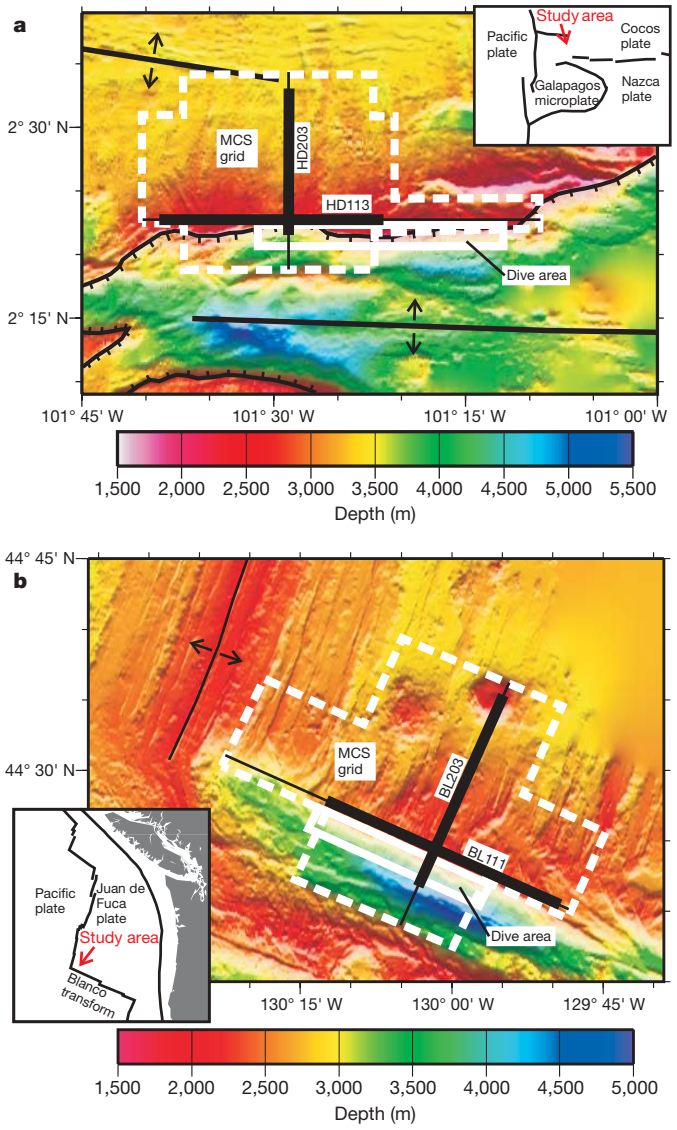


Figure 1 | Study areas and bathymetry maps. **a**, HDR study area; **b**, BTF study area. White dashed lines outline the multichannel seismic (MCS) data grid, and white solid lines enclose the dive area. Heavy solid lines indicate the locations of the MCS profiles shown in Fig. 3. Insets show the regional tectonic setting.

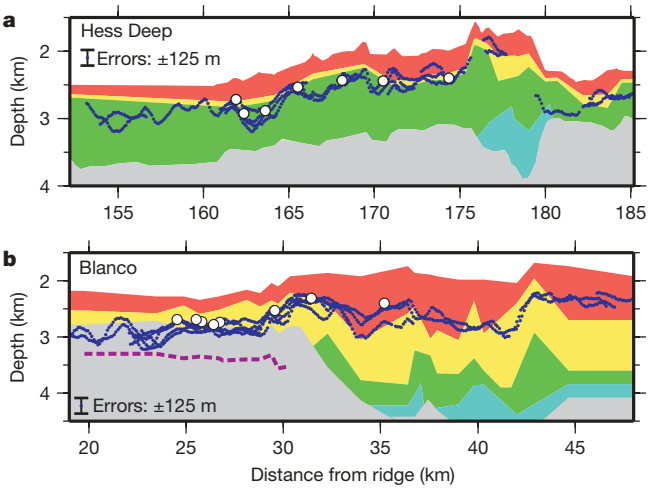


Figure 2 | Geologic cross-sections and layer 2A event picks. **a**, HDR study area; **b**, BTF study area. Shaded regions indicate the geologic units: lavas (red); transition zone between lavas and dykes (yellow); sheeted dyke complex (green); gabbroic unit or transition zone between dykes and gabbros (cyan); and talus (grey). Layer 2A event picks from profiles within 2.5 km of the scarps, converted to depth and projected onto the cross-sections, are shown in blue. Layer 2A/2B boundaries from velocity analyses of common depth point (CDP) supergathers (see Supplementary Information for more details and error bar explanation) are projected onto the cross-section and shown with white circles. In **a**, layer 2A event picks from profiles HD116, HD115, HD114, HD113 and HD112 are displayed. In **b**, layer 2A event picks from profiles BL113, BL112, BL111, BL110 and BL109 are displayed. Dashed magenta line indicates the depth of the top of the sheeted dykes interpreted from magnetic profiles[18].

Table 1 | Comparison of observations between the HDR and the BTF

Observation	HDR	BTF
Mean TWTT to layer 2A event	0.37 s	0.36 s
Mean layer 2A velocity*	2.60 km s^{-1}	2.70 km s^{-1}
Mean layer 2A thickness	0.48 km	0.49 km
Mean velocity at top of layer 2B†	4.8 km s^{-1}‡, 5.15 km s^{-1}§	4.3 km s^{-1}‡, 5.0 km s^{-1}§
Mean lava thickness	0.30 km	0.45 km
Mean transition zone thickness	0.15 km	0.70 km
Mean depth below sea floor‖	0.45 km	1.15 km
Layer 2A depth compared to geology	Near top of dykes	Generally in transition zone

* From travel-time modelling of selected supergathers (see Supplementary Information for more details).
† From streamer refractions.
‡ Near scarp.
§ Elsewhere.
‖ Top of sheeted dykes.

in that unit; however, at the HDR, the seismic layer 2A/2B boundary is commonly located within the dykes (Fig. 2a) and might suggest an alteration boundary in the dyke unit.

It is possible that the nature of the layer 2A/2B boundary is dependent on spreading rate. For fast-spreading crust such as the HDR, the uniform accretionary system may result in the layer 2A/2B boundary corresponding to the geologic boundary between lava and dykes, whereas for intermediate-spreading crust such as the BTF, more varied unit thickness and fracturing may result in the layer 2A/2B boundary corresponding to an upper crustal alteration boundary. However, we argue that the similarity in the seismic properties of the layer 2A/2B boundary (Table 1 and Supplementary Information) at the HDR and the BTF, and the similarity in the visual appearance of the seismic profiles (Fig. 3) at these two sites, support a common origin for the layer 2A event at the two study areas. The increase in seismic velocity at ~400–600 m depth must be associated with a change in physical properties, but the geologic observations indicate that this change occurs in different crustal units in the different crustal sections examined. Increasing confining pressure with depth will close cracks and increase seismic velocity, but will only result in a gradual velocity gradient with depth and cannot be solely responsible for the high-gradient region observed at the base of layer 2A. Our preferred explanation for the nature of the layer 2A/2B boundary is that it is associated with an alteration front that may occur within either the lavas or dykes, depending on the thickness of the lava layer. Below the front, cracks are healed by hydrothermal metamorphism, porosities are decreased, and seismic velocities are increased. Crack closure with depth owing to confining pressure will assist this process, and there may be a crack thickness threshold near 400–600 m depth at which cracks are easily sealed with alteration products.

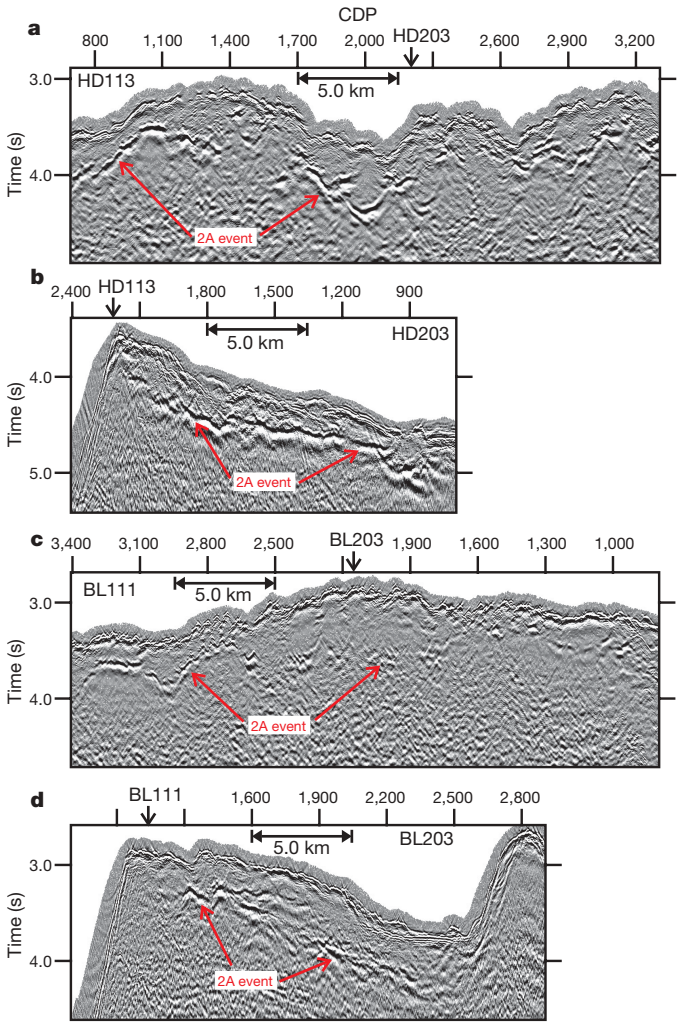


Figure 3 | Layer 2A event imaged on seismic reflection data. Locations of profiles are shown in Fig. 1. Arrows indicate location of crossing lines. HDR profiles: **a**, HD113; **b**, HD203. BTF profiles: **c**, BL111; **d**, BL203.

A zeolite alteration zone is observed within the lavas at DSDP Site 504B[25] and would be the likely candidate for the source of the layer 2A/2B boundary at the BTF. Indeed, abundant fractures filled with zeolites have been observed in some lava samples at the BTF[17]. At DSDP Site 504B[23] and ODP Site 1256D[24], strongly hydrothermally altered rocks (greenschist facies) appear 25–50 m below the top of the transition zone, and porosities decrease rapidly downward from the top of the transition zone at the former site[23] and at the tops of the sheeted dykes at the latter site[24]. These physical property changes near the top of the dykes[15,26] are consistent with the position of the layer 2A/2B boundary at the HDR.

A primary observation of the layer 2A/2B boundary is that it approximately doubles in thickness within 1–2 km of the ridge crest for fast-spreading crust; this is commonly explained by a doubling in thickness of the lava layer accompanied by subsidence of the sheeted dykes[2,3,27–29]. Karson and Christeson[30] argue that the rock material corresponding to seismic layer 2A changes as it evolves along a spreading flow line. In their model, layer 2A in zero-age crust corresponds mostly to weakly fractured sheeted dykes and overlying lavas. Off axis, fracturing from the top down thickens the low-velocity layer. At depth, hydrothermal healing of fractures and dyke intrusions may convert the base of layer 2A to higher-velocity layer 2B. The net result of this process would be the observed off-axis thickening of layer 2A without requiring substantial thickening of the lava unit. Thus, the layer 2A/2B boundary in off-axis crust would be an alteration front occurring within the lavas, the dykes, or the transition zone between these units.

In summary, our study indicates that the seismic layer 2A/2B boundary does not universally correlate with the structural boundary between lavas and dykes. It is more likely that the primary control on the depth of the layer 2A/2B boundary is crack closure enhanced by hydrothermal alteration and sealing. Further mapping of alteration distribution within the upper oceanic crust is required to confirm this hypothesis. Numerous studies have mapped the layer 2A/2B

boundary over extensive regions of young oceanic crust; if this seismic boundary does correspond to an alteration boundary, then existing seismic data sets can provide new constraints on the porosity structure of the upper oceanic crust.

Received 25 July; accepted 5 December 2006.

1. Herron, T. J. Lava flow layer — East Pacific Rise. *Geophys. Res. Lett.* **9,** 17–20 (1982).
2. Christeson, G. L., Purdy, G. M. & Fryer, G. J. Structure of young upper crust at the East Pacific Rise near 9°30′N. *Geophys. Res. Lett.* **19,** 1045–1048 (1992).
3. Harding, A. J., Kent, G. M. & Orcutt, J. A. A multichannel seismic investigation of upper crustal structure at 9°N on the East Pacific Rise: Implications for crustal accretion. *J. Geophys. Res.* **98,** 13925–13944 (1993).
4. Rohr, K. M. M., Milkereit, B. & Yorath, C. J. Asymmetric deep crustal structure across the Juan de Fuca Ridge. *Geology* **16,** 533–537 (1988).
5. Harding, A. J. *et al.* Structure of young oceanic crust at 13°N on the East Pacific Rise from expanding spread profiles. *J. Geophys. Res.* **94,** 12163–12196 (1989).
6. Vera, E. E. *et al.* The structure of 0- to 0.2-m.y.-old oceanic crust at 9°N on the East Pacific Rise from expanded spread profiles. *J. Geophys. Res.* **95,** 15529–15556 (1990).
7. Hooft, E. E. E., Schouten, H. & Detrick, R. S. Constraining crustal emplacement processes from the variation in seismic Layer 2A thickness at the East Pacific Rise. *Earth Planet. Sci. Lett.* **142,** 289–309 (1996).
8. Buck, W. R., Carbotte, S. M. & Mutter, C. Controls on extrusion at mid-ocean ridges. *Geology* **25,** 935–938 (1997).
9. Carbotte, S., Mutter, C., Mutter, J. & Ponce-Correa, G. Influence of magma supply and spreading rate on crustal magma bodies and emplacement of the extrusive layer: Insights from the East Pacific Rise at lat 16°N. *Geology* **26,** 455–458 (1998).
10. Schouten, H., Tivey, M. A., Fornari, D. J. & Cochran, J. R. Central anomaly magnetization high: Constraints on the volcanic construction and architecture of seismic layer 2A at a fast-spreading mid-ocean ridge, the EPR at 9°30′-50′N. *Earth Planet. Sci. Lett.* **169,** 37–50 (1999).
11. Searle, R. & Francheteau, J. Morphology and tectonics of the Galapagos triple junction. *Mar. Geophys. Res.* **8,** 95–129 (1986).
12. Lonsdale, P. Structural pattern of the Galapagos microplate and evolution of the Galapagos triple junctions. *J. Geophys. Res.* **93,** 13551–13574 (1988).
13. Francheteau, J. *et al.* 1 Ma East Pacific Rise oceanic crust and uppermost mantle exposed by rifting in Hess Deep (equatorial Pacific Ocean). *Earth Planet. Sci. Lett.* **101,** 281–295 (1990).
14. Francheteau, J. *et al.* Dyke complex of the East Pacific Rise exposed in the walls of Hess Deep and the structure of the upper oceanic crust. *Earth Planet. Sci. Lett.* **111,** 109–121 (1992).
15. Karson, J. A. *et al.* Structure of uppermost fast-spread oceanic crust exposed at the Hess Deep Rift: Implications for subaxial processes at the East Pacific Rise. *Geochem. Geophys. Geosyst.* **3,** doi:10.1029/2001GC000155 (2002).
16. Delaney, J. R., Johnson, H. P. & Karsten, J. L. The Juan de Fuca Ridge - hot spot - propagating rift system: New tectonic, geochemical, and magnetic data. *J. Geophys. Res.* **86,** 11747–11750 (1981).
17. Juteau, T. *et al.* A submersible study in the Western Blanco Fracture Zone, N.E. Pacific: Structure and evolution during the last 1.6 Ma. *Mar. Geophys. Res.* **17,** 399–430 (1995).
18. Tivey, M. A. *et al.* Direct measurement of magnetic reversal polarity boundaries in a cross-section of oceanic crust. *Geophys. Res. Lett.* **25,** 3631–3634 (1998).
19. Karson, J. A., Tivey, M. A. & Delaney, J. R. Internal structure of uppermost oceanic crust along the Western Blanco Transform Scarp: Implications for subaxial accretion and deformation at the Juan de Fuca Ridge. *J. Geophys. Res.* **107,** doi:10.1029/2000JB000051 (2002).
20. Christeson, G. L., Purdy, G. M. & Fryer, G. J. Seismic constraints on shallow crustal emplacement processes at the fast-spreading East Pacific Rise. *J. Geophys. Res.* **99,** 17957–17973 (1994).
21. Vera, E. E. & Diebold, J. B. Seismic imaging of oceanic layer 2A between 9°30′N and 10°N on the East Pacific Rise from two-ship wide-aperture profiles. *J. Geophys. Res.* **99,** 3031–3041 (1994).
22. Canales, J. P. *et al.* Upper crustal structure and axial topography at intermediate spreading ridges: Seismic constraints from the southern Juan de Fuca Ridge. *J. Geophys. Res.* **110,** doi:10.1029/2005JB003630 (2005).
23. Alt, J. C. *et al.* Hydrothermal alteration of a section of upper oceanic crust in the eastern Equatorial Pacific; a synthesis of results from Site 504 (DSDP legs 69–70, and 83, and ODP legs 111, 137, 140, and 148). *Proc. ODP Sci. Res.* **148,** 417–434 (1996).
24. Expedition 309 and 312 Scientists.. Superfast spreading rate crust 3: a complete in situ section of upper oceanic crust formed at a superfast spreading rate. *IODP Prelim. Rep.* **312,** 45–50 (2006).
25. Alt, J. C., Honnorez, J., Laverne, C. & Emmermann, R. Hydrothermal alteration of a 1 km section through the upper oceanic crust, Deep Sea Drilling Project Hole 504B: Mineralogy, chemistry, and evolution of seawater-basalt interactions. *J. Geophys. Res.* **91,** 10309–10335 (1986).
26. Gillis, K. M. Controls on hydrothermal alteration in a section of fast-spreading oceanic crust. *Earth Planet. Sci. Lett.* **134,** 473–489 (1995).
27. Kent, G. M. *et al.* Uniform accretion of oceanic crust south of the Garrett transform at 14°15′S on the East Pacific Rise. *J. Geophys. Res.* **99,** 9097–9116 (1994).
28. Toomey, D. R., Purdy, G. M., Solomon, S. C. & Wilcock, W. S. D. The three-dimensional seismic velocity structure of the East Pacific Rise near latitude 9° 30′ N. *Nature* **347,** 639–645 (1990).
29. Caress, D. W., Burnett, M. S. & Orcutt, J. A. Tomographic image of the axial low-velocity zone at 12°50′N on the East Pacific Rise. *J. Geophys. Res.* **97,** 9243–9263 (1992).
30. Karson, J. A. & Christeson, G. L. in *Heterogeneity in the Crust and Upper Mantle: Nature, Scaling and Seismic Properties* (eds Goff, J. & Holliger, K.) 99–129 (Kluwer Academic, New York, 2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to the captain, crew and science parties of cruises EW0305 and EW0410 of the R/V *Maurice Ewing* for their assistance. We also thank G. Kent for comments on the manuscript. This work was supported by the National Science Foundation. This is a UTIG contribution.

Author Contributions G.L.C. and K.D.M. processed and interpreted the seismic data. All authors contributed equally to the integration of the seismic data and geologic observations.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to G.L.C. (gail@ig.utexas.edu).

LETTERS

An arid-adapted middle Pleistocene vertebrate fauna from south-central Australia

Gavin J. Prideaux[1], John A. Long[1,2], Linda K. Ayliffe[3,4], John C. Hellstrom[5], Brad Pillans[4], Walter E. Boles[6], Mark N. Hutchinson[7], Richard G. Roberts[8], Matthew L. Cupper[5], Lee J. Arnold[8], Paul D. Devine[9] & Natalie M. Warburton[1]

How well the ecology, zoogeography and evolution of modern biotas is understood depends substantially on knowledge of the Pleistocene[1,2]. Australia has one of the most distinctive, but least understood, Pleistocene faunas. Records from the western half of the continent are especially rare[3]. Here we report on a diverse and exceptionally well preserved middle Pleistocene vertebrate assemblage from caves beneath the arid, treeless Nullarbor plain of south-central Australia. Many taxa are represented by whole skeletons, which together serve as a template for identifying fragmentary, hitherto indeterminate, remains collected previously from Pleistocene sites across southern Australia. A remarkable eight of the 23 Nullarbor kangaroos are new, including two tree-kangaroos. The diverse herbivore assemblage implies substantially greater floristic diversity than that of the modern shrub steppe, but all other faunal and stable-isotope data indicate that the climate was very similar to today. Because the 21 Nullarbor species that did not survive the Pleistocene were well adapted to dry conditions, climate change (specifically, increased aridity) is unlikely to have been significant in their extinction.

The Nullarbor plain is a vast treeless expanse of chenopod shrub steppe[4] at the centre of the 240,000 km^2 onshore portion of the Eucla basin[5] (Fig. 1). In 2002, a speleological team discovered vertebrate fossils in three Nullarbor caves. Here named the Thylacoleo caves, these large collapse chambers preserve the most diverse Pleistocene fauna yet recovered from the western half of Australia. No vertebrate remains older than late Pleistocene[6] have previously been detailed from south-central Australia. Their discovery provides a long-awaited historical perspective on the region that today acts as a barrier to the east–west dispersal of many taxa[7,8]. Fossil preservation in the Thylacoleo caves is without precedent in Australia. Several new and previously incompletely known species are represented by whole skeletons (Fig. 2).

U/Pb ages of 4.1–3.8 Myr ago[9] on *in situ* stalagmites in Leaena's Breath cave (LBC) indicate that this cave formed before the mid-Pliocene. LBC and Flightstar cave (FSC) are each entered through a solution pipe in the chamber roof, with pitches of about 20 m to the floors below. The preservation and distribution of cave-surface fossils (decreasing concentration away from entrances) suggests that Pleistocene animals fell or flew down these same pipes before their blocking. An obstructed pipe 8 m away from the present collapse entrance of Last Tree cave (LTC) was the likely Pleistocene opening. Although most animals were evidently pitfall victims, sediments in FSC and LBC preserve a high proportion of small bones that may derive from regurgitated pellets of roosting owls (*Tyto* species) or kestrels (*Falco cenchroides*), which are present in the fauna. Fossils occur within sediment formed from limestone breakdown, atop or wedged between boulders, and atop or buried in silty clay infill sediments (see Supplementary Information).

Three dating techniques and reconstruction of deposition processes permit the recognition of at least two intervals of faunal

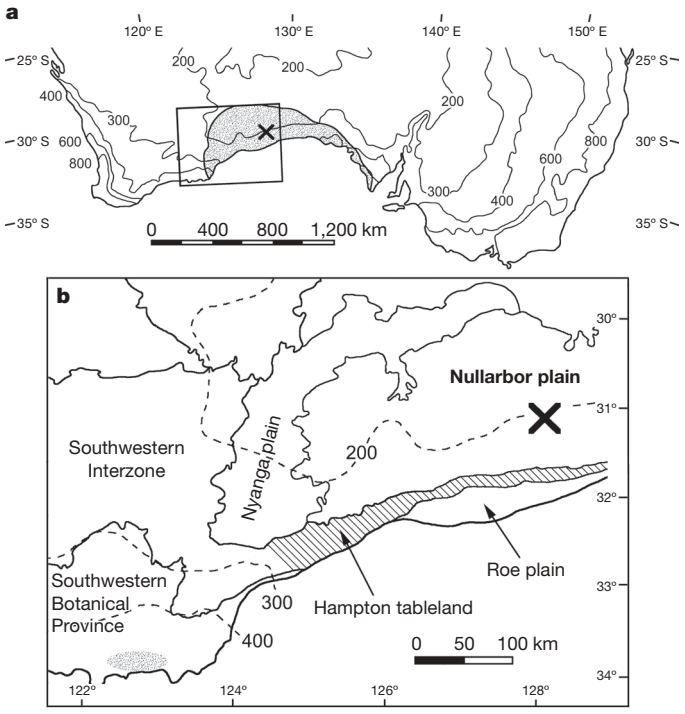


Figure 1 | Location of the Nullarbor plain and Thylacoleo caves. a, Southern mainland Australia, showing the Eucla basin (stippled) and mean annual rainfall isohyets (mm). **b**, Botanical regions of the western Eucla basin and surrounds[4,5]. Nullarbor plain, unwooded shrub steppe; Hampton tableland and Roe/Nyanga plains, wooded shrub steppe and shrubland; Southwestern Botanical Province and Southwestern Interzone, shrubland, heath and mixed woodland. Dashed lines denote rainfall[19]. The stippled area marks the nearest modern occurrences of *Macropus eugenii* and *Egernia kingii*. The hatched area is Hampton tableland, and the cross represents the location of the Thylacoleo caves.

[1]Department of Earth and Planetary Sciences, Western Australian Museum, Perth, Western Australia 6000, Australia. [2]Museum Victoria, PO Box 666, Melbourne, Victoria 3001, Australia. [3]Department of Geology and Geophysics, University of Utah, Salt Lake City, Utah 84112, USA. [4]Research School of Earth Sciences, Australian National University, Canberra, Australian Capital Territory 0200, Australia. [5]School of Earth Sciences, University of Melbourne, Melbourne, Victoria 3010, Australia. [6]Terrestrial Zoology, Australian Museum, Sydney, New South Wales 2010, Australia. [7]Herpetology Section, South Australian Museum, Adelaide, South Australia 5000, Australia. [8]GeoQuEST Research Centre, School of Earth and Environmental Sciences, University of Wollongong, Wollongong, New South Wales 2522, Australia. [9]Speleological Research Group Western Australia, PO Box 1611, East Victoria Park, Western Australia 6981, Australia.

accumulation in LBC. Palaeomagnetic samples from the infill sediment reveal a magnetic reversal between 0.5 and 0.7 m depth (see Supplementary Information). The lower sediments accumulated during the Matuyama Reversed Chron (more than 780 kyr ago). Deposition ceased when the entrance pipe became blocked. The upper sediments accumulated, after reopening of the pipe, during the Brunhes Normal Chron (less than 780 kyr ago). Optical dating of quartz grains provides a minimum age of 195 ± 15 kyr (mean ± s.e.m.) for entry of the latter sediments into LBC. The flood responsible for a surface palaeochannel (containing partial skeletons and crania) may have induced blockage of the pipe, until its recent reopening.

In places, the upper LBC sediment overlies coralline calcite speleothems. $^{234}U/^{230}Th$ ages of 407 ± 17, 394 ± 15 and 392 ± 24 kyr have been obtained from coralline samples, the latter encrusting a cave-surface fossil. These constrain the age of the upper sediment and its fossils to some time between 400 and 200 kyr ago (see Supplementary Information). Marked similarity in preservation suggests that most if not all of the cave-surface material exceeds 400 kyr in age. Optical dating of surface sediments in FSC and LTC provides minimum ages for fossils of 230 ± 27 kyr and 101 ± 17 kyr, respectively.

Sixty-nine vertebrate and one gastropod species have been identified in the Thylacoleo caves (see Supplementary Information). This includes 23 kangaroo species, eight of which are undescribed (Table 1). Several are represented by complete crania and/or skeletons, facilitating the identification of three of these 'new' species among fragmentary, indeterminate material collected previously from Pleistocene sites in semi-arid southern Australia (Table 1). Herbivore diversity is similar to that of other Pleistocene assemblages[10–12], but the fauna from the Thylacoleo caves has proportionally fewer arboreal folivores and fungivores, and a larger number of mixed feeders and grazers, similar to the Pleistocene fauna of inland Australia (see Supplementary Information). This suggests a dry, relatively open environment.

Among the 20 mammals in the fauna that survived the Pleistocene, only *Macropus eugenii* and *Dasycercus* sp. cf. *blythi* have no regional late Holocene records[13,14] (Fig. 1b). With the exception of a specimen tentatively identified as *Egernia kingii* (a species now restricted to the Southwestern Botanical Province), the remaining lizards are common locally or in peripheral open woodland or shrubland habitats (http://www.amonline.net.au/herpetology/research/index.htm) (Fig. 1b). Overall, the presence of the lizards indicates dry conditions, but with a more diverse vegetation than now exists on the Nullarbor plain. The presence of birds rules out closed woodland, but they are otherwise generalized in their habitat preferences (see Supplementary Information). However, the presence of two parrots suggests the presence of hollow-forming trees useful for nesting. The gastropod, *Pupoides adelaidae*, ranges widely today through semi-arid and arid southern Australia, including the Nullarbor plain[15].

Stable carbon-isotope and oxygen-isotope ratios in herbivore teeth reflect the isotope contents of diet vegetation and ingested waters[16,17], which are in turn influenced by climate[18]. We analysed 59 enamel samples from 13 kangaroo and 1 wombat species in the Thylacoleo caves fauna, and compared these with a data set based on modern grazing kangaroo and wombat specimens (Fig. 3) from the winter, uniform and arid/non-seasonal rainfall zones of southern Australia[3,19]. Low summer rainfall in southern Australia determines the predominance of C_3 grasses in areas of low, medium and high rainfall (Fig. 3a). In the Nullarbor region, modern samples from the Hampton tableland (Fig. 1b) lie within the C_3 range, whereas those from the northern Nullarbor plain give mixed (C_3 and C_4) signatures (Fig. 3b). $\delta^{13}C$ values for Pleistocene herbivores from the Thylacoleo

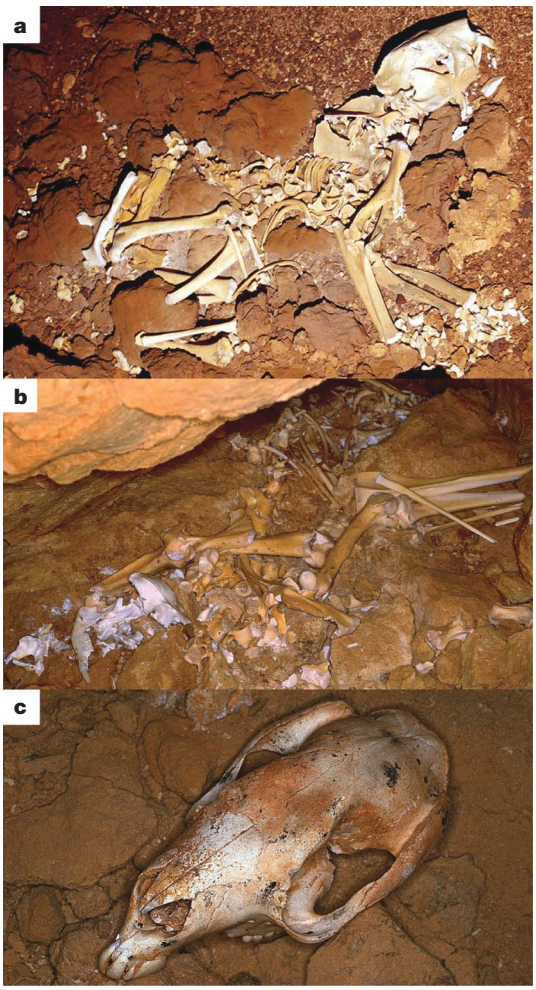


Figure 2 | Fossils from the Thylacoleo caves. a, Complete skeleton of *Thylacoleo carnifex* (WAM 02.7.1). **b**, Complete skeleton of *Baringa* sp. nov. 1 (WAM 02.7.17). This unusual kangaroo is the most common marsupial in the fauna, and bore high-crowned incisors and enlarged tuberosities above its eye orbits. **c**, Cranium of *Sthenurus andersoni* (WAM 03.5.5).

Table 1 | Species in the Thylacoleo caves fauna that did not survive the Pleistocene

Species	Rainfall range (mm)		Body
	Lower	Upper	mass (kg)
*Leipoa gallinacea**	200	690	10
*Phascolonus gigas**	150	800	200
*Thylacoleo carnifex**	150	1,200	80
Baringa sp. nov. 1	–	260	30
Baringa sp. nov. 2	–	–	20
Baringa sp. nov. 3	–	–	30
Bohra sp. nov. 1	–	500	25
Bohra sp. nov. 2	–	–	25
*Congruus kitcheneri**	–	1,000	40
Congruus sp. nov. 1	–	250	50
Congruus sp. nov. 2	–	–	50
*Macropus ferragus**	150	800	150
Macropus sp. nov.	150	1,000	35
*Metasthenurus newtonae**	260	1,200	55
*Procoptodon goliah**	100	800	200
*'Procoptodon' browneorum**	250	1000	60
'Procoptodon' williamsi	150	600	120
*Protemnodon brehus**	150	1,000	100
*Protemnodon roechus**	150	800	120
*Sthenurus andersoni**	100	1,000	50
*Sthenurus tindalei**	100	300	100
Minimum	100	260	10
Maximum	260	1,200	200

Estimated rainfall bounds (upper and lower) are based on the modern mean annual rainfall[19] across Pleistocene ranges[3]. The four species with neither bound are unknown outside of the Thylacoleo caves; those lacking a lower bound are known from only one other site. Body masses are from ref. 29 or have been estimated from a comparison with closest-sized living species. Species marked by an asterisk have late Pleistocene records.

caves (−13.3 to −7.2‰; $n = 41$) cluster between the modern Hampton tableland and northern Nullarbor samples (Fig. 3b).

Modern samples from warmer, drier areas of southern Australia yield enamel $\delta^{18}O$ values that are more positive than those from cooler, wetter areas (Fig. 3a). $\delta^{18}O$ values from the Hampton tableland (mean annual rainfall 240–270 mm) herbivores range from −4.4 to 4.2‰ ($n = 71$), whereas those from the northern Nullarbor (180 mm) range from 2.0 to 8.4‰ ($n = 5$) (Fig. 3b). The Pleistocene herbivores yield $\delta^{18}O$ values (−0.1 to 7.6‰; $n = 40$) intermediate between those of the modern samples (Fig. 3b), despite slight differences between sites that may or may not (given low sample sizes) reflect slight variations in prevailing climate due to site age differences. Nevertheless, the overriding climatic implication of the isotope data is manifest: effective precipitation and seasonality during the intervals over which the Pleistocene fauna accumulated were similar to those of the present, in which an annual mean of about 200 mm falls in a largely non-seasonal, but slightly winter-biased, pattern[19].

Thirty-one of the 36 Thylacoleo caves species that survived the Pleistocene have recent records in the region[13,14]. Estimated from modern tolerances (see Supplementary Information), all species except *Macropus eugenii* and *Egernia* sp. cf. *kingii* could have coexisted within a 230–250-mm mean annual rainfall range. Factors controlling the modern distribution of *M. eugenii* and *E. kingii* are uncertain, but it is likely that shrub diversity, not rainfall, is a key determinant. Modern rainfall across the former ranges[3] of the 21 species that became extinct before the Holocene (excluding the four known only from the Thylacoleo caves) suggests that they could have cohabited an area with a rainfall of about 260 mm (Table 1). Together with the modern faunal and isotope data, these indicate a predominantly arid climate with a palaeorainfall range of 230–260 mm.

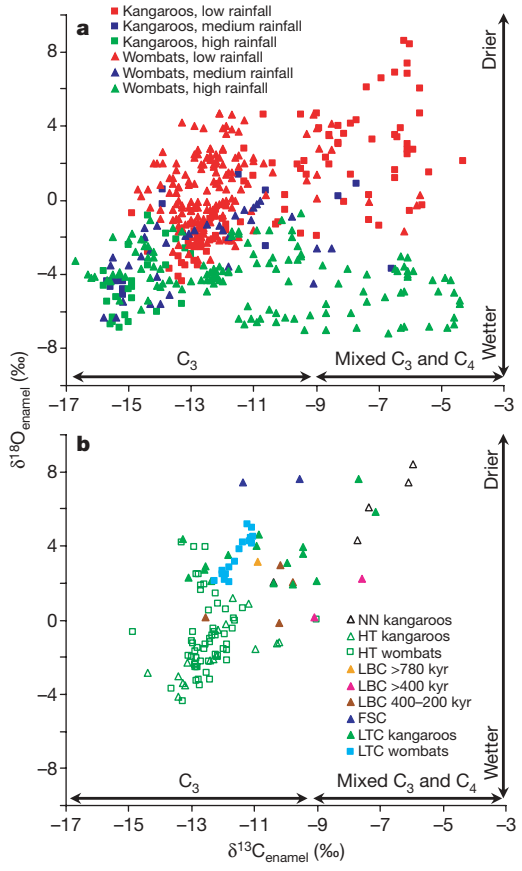


Figure 3 | Stable carbon-isotope and oxygen-isotope values ($\delta^{13}C$ and $\delta^{18}O$) from the enamel of kangaroos and wombats. a, Modern samples from the winter, uniform and arid/non-seasonal rainfall zones of southern Australia[3,19]. Low rainfall, less than 300 mm; medium rainfall, 350–550 mm; high rainfall, more than 600 mm. **b**, Isotope ratios of fossil samples from the Thylacoleo caves (LBC, FSC, LTC) compared with those of modern samples from the Hampton tableland (HT) and northern Nullarbor plain (NN).

Cave sediments are too oxidized to preserve pollen, but the high diversity of herbivores reflects a profoundly different vegetation structure from that of the modern chenopod steppe. No species exemplify this more than the tree-kangaroos (*Bohra*), whose modern relatives (*Dendrolagus*) inhabit rainforest. In stark contrast, however, *Bohra* species were not restricted to well-wooded habitats[3], and probably represent the hitherto 'missing' arboreal folivores of the semi-arid and arid woodlands of late Cenozoic Australia. The varied morphologies and broad size range (4–200 kg) of the 20 terrestrial browsing and/or grazing marsupials (Table 1; Supplementary Information) indicate a scleromorphic mosaic of woodland and shrubland incorporating a higher proportion of plants with palatable leaves and fleshy fruits (for example, species of Myoporaceae, Santalaceae, Pittosporaceae and Loranthaceae), such as those now largely confined to remnant stands on the Nullarbor periphery[4,20] (Fig. 1b; see Supplementary Information). Loss of a similar range of fire-sensitive plants, and their replacement by fire-adapted mulga woodland and spinifex grassland, as a result of landscape burning by humans has been advanced as a cause of the extinction in central Australia of the giant flightless bird *Genyornis*[21,22].

By establishing that the Nullarbor Pleistocene fauna was adapted to dry conditions, extinction hypotheses invoking megafaunal susceptibility to aridity[23,24] become untenable. Even during the driest times, ranges of the less arid-adapted species could simply have shifted more peripherally. Indeed, 12 of the 13 described Nullarbor megafaunal species survived into the late Pleistocene elsewhere, and were common in drier areas[3,11,25] (Table 1). It follows that climate change alone is unlikely to have precipitated the demise of this remarkable fauna, particularly in view of other evidence that Australian megafaunal species were resilient to glacial–interglacial cycling[10]. We argue that increased wildfires in the Nullarbor region best explain the conversion of a floristically diverse plant community into the modern, fire-resistant, chenopod shrub steppe. Although capable of supporting abundant vegetation, limestone substrates are inherently infertile, with nutrients held in thin upper soil horizons highly susceptible in arid areas to pyrogenic denudation[26]. Our data do not directly explain the timing of extinctions, but it is significant that the general extinction pattern (the loss of most larger herbivores and *Thylacoleo*) is identical to that witnessed in all southern Australian climatic zones[10,12,25]. Most southern species of megafauna were evidently extinct by or soon after 40 kyr ago[6,12,25], at about the time humans reached the south-central coast[27].

METHODS

Palaeontology. Fossils were photographed *in situ* and locations were recorded relative to cave survey points. All specimens were extremely fragile and required hardening with polyvinyl butyrate dissolved in 100% ethanol before being wrapped and removed from sites in hard plastic cases. On all occasions, great care was taken to preserve element associations. Sediments were removed from the LBC test excavation in 10-cm levels and wet-screened for small vertebrate remains. All specimens are lodged in the Western Australian Museum, Perth.

Geochronology. $^{234}U/^{230}Th$ dating was conducted on solid pieces of calcite weighing 3–30 mg, cut from the speleothem samples with a dental drill. Multiple subsamples were taken from each coralline crust to assess the degree of open-system behaviour for these samples. Eight oriented monoliths ($10 \times 10 \times 5\ cm^3$) were carefully excavated from a $0.5\ m^2 \times 1.3$ m-deep test pit in sediments in LBC to assess palaeomagnetism. Standard palaeomagnetic samples ($2 \times 2 \times 2\ cm^3$) were prepared from each monolith. Three samples from each monolith were measured on a 2G Enterprises three-axis cryogenic magnetometer. Stepwise alternating field demagnetization (peak field 140 mT) was undertaken to isolate characteristic remanent magnetizations, which were determined by principal components analysis. Optical dating provides an estimate of time elapsed since luminescent minerals, such as quartz, were last exposed to sunlight[28]. In this study, the event being dated was the time of entry of sediment grains into the caves. Optical ages for buried quartz grains were calculated from the burial dose (estimated from the optically stimulated luminescence signal) divided by the dose rate due to ionizing radiation.

Stable isotopes. Samples of powdered enamel from fourth molars were subjected to 15-min treatments with 3% hydrogen peroxide followed by 0.1 M acetic acid with several intervening rinses with demineralized water. Samples of 0.5–1.0 mg were reacted at 90 °C in a common acid bath. Evolved gases were cryogenically purified of H_2O and transferred to a microvolume cold finger before analysis through the dual inlet of a Finnigan MAT 252 mass spectrometer. Isotope results were standardized to the Pee Dee Belemnite scale by in-run comparison of enamel standards calibrated against NBS-19. $\delta^{13}C$, $\delta^{18}O = (R_{sample}/R_{standard} - 1) \times 1{,}000$, where R is the $^{13}C/^{12}C$ or $^{18}O/^{16}O$ ratio.

Received 25 August; accepted 21 November 2006.

1. FAUNMAP Working Group. Spatial response of mammals to Late Quaternary environmental fluctuations. *Science* **272,** 1601–1606 (1996).
2. Lister, A. M. The impact of Quaternary Ice Ages on mammalian evolution. *Phil. Trans. R. Soc. Lond. B* **359,** 221–241 (2004).
3. Prideaux, G. J. in *Encyclopedia of Quaternary Science* (ed. Elias, S.) 1518–1537 (Elsevier, Oxford, 2006).
4. Beard, J. S. *Vegetation Survey of Western Australia: Nullarbor. 1:1,000,000 Vegetation Series. Explanatory Notes to Sheet 4* (University of Western Australia, Perth, 1975).
5. Lowry, D. C. Geology of the Western Australian part of the Eucla Basin. *Geol. Surv. West. Aust. Bull.* **122,** 1–201 (1970).
6. Lundelius, E. L. & Turnbull, W. D. The mammalian fauna of Madura Cave, Western Australia. Part VII: Macropodidae: Sthenurinae, Macropodinae, with a review of the marsupial portion of the fauna. *Fieldiana Geol.* NS **17,** 1–71 (1989).
7. Burbidge, N. T. The phytogeography of the Australian region. *Aust. J. Bot.* **8,** 75–212 (1960).
8. Keast, A. (ed.) *Ecological Biogeography of Australia* (W. Junk, The Hague, 1981).
9. Woodhead, J. *et al.* U–Pb geochronology of speleothems by MC-ICPMS. *Quat. Geochronol.* **1,** 208–221 (2006).
10. Prideaux, G. J. *et al.* Mammalian responses to Pleistocene climate change in southeastern Australia. *Geology* **35,** 33–36 (2007).
11. Prideaux, G. J. *Systematics and Evolution of the Sthenurine Kangaroos* (Univ. Calif. Publ. Geol. Sci. no. 146, 2004).
12. Prideaux, G. J., Gully, G. A., Ayliffe, L. K., Bird, M. I. & Roberts, R. G. Tight Entrance Cave, southwestern Australia: a late Pleistocene vertebrate deposit spanning more than 180 ka. *J. Vertebr. Paleontol.* **20,** 62A–63A (2000).
13. Baynes, A. in *A Biological Survey of the Nullarbor Region, South and Western Australia in 1984* (eds McKenzie, N. L. & Robinson, A. C.) 139–152 (Department of Environment and Planning, Adelaide, 1987).
14. Strahan, R. (ed.) *The Mammals of Australia* (Reed New Holland, Sydney, 1995).
15. Solem, A. Pupilloid land snails from the south and mid-west coasts of Australia. *J. Malacol. Soc. Aust.* **7,** 95–124 (1986).
16. Passey, B. H. *et al.* Carbon isotope fractionation between diet, breath CO_2, and bioapatite in different mammals. *J. Archaeol. Sci.* **32,** 1459–1470 (2005).
17. Cerling, T. E. *et al.* The reaction progress variable in stable isotope studies of biological tissue turnover. *Oecologia* (in the press).
18. Koch, P. L. Isotopic reconstruction of past continental environments. *Annu. Rev. Earth Planet. Sci.* **26,** 573–613 (1998).
19. Bureau of Meteorology. *Climate of Australia* (Australian Government Publishing Service, Canberra, 1989).
20. Mitchell, A. A. & Wilcox, D. G. *Arid Shrubland Plants of Western Australia* (Univ. of Western Australia Press, Perth, 1998).
21. Miller, G. H. *et al.* Ecosystem collapse in Pleistocene Australia and a human role in megafaunal extinction. *Science* **309,** 287–290 (2005).
22. Murray, P. F. & Vickers-Rich, P. *Magnificent Mihirungs: The Colossal Flightless Birds of the Australian Dreamtime* (Univ. of Indiana Press, Bloomington, Indiana, 2004).
23. Horton, D. R. in *Quaternary Extinctions: A Prehistoric Revolution* (eds Martin, P. S. & Klein, R. G.) 639–680 (Univ. of Arizona, Tucson, 1984).
24. Wroe, S. & Field, J. A review of the evidence for a human role in the extinction of Australian megafauna and an alternative interpretation. *Quat. Sci. Rev.* **25,** 2692–2703 (2006).
25. Roberts, R. G. *et al.* New ages for the last Australian megafauna: continent-wide extinction about 46,000 years ago. *Science* **292,** 1888–1892 (2001).
26. McKenzie, N. J., Jacquier, D., Isbell, R. F. & Brown, K. *Australian Soils and Landscapes: An Illustrated Compendium* (CSIRO Publishing, Melbourne, 2004).
27. Roberts, R. G. *et al.* Preliminary luminescence dates for archaeological sediments on the Nullarbor Plain, South Australia. *Aust. Archaeol.* **42,** 7–16 (1996).
28. Lian, O. B. & Roberts, R. G. Dating the Quaternary: progress in luminescence dating of sediments. *Quat. Sci. Rev.* **25,** 2449–2468 (2006).
29. Johnson, C. N. & Prideaux, G. J. Extinctions of herbivorous mammals in Australia's late Pleistocene in relation to their feeding ecology: no evidence for environmental change as cause of extinction. *Aust. Ecol.* **29,** 553–557 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank P. Ackroyd, K. Boland, R. Gibbons, G. MacLucas, J. MacLucas and E. Taylor for bringing their discoveries to our attention, and for cave mapping and field assistance; C. Bryce, G. Deacon, L. Hatcher, D. Megirian and M. Norton for field assistance; G. Kendrick and A. Baynes for identifying gastropod and rodent fossils, respectively; and P. Latz, J. Magee, D. Megirian, G. Miller and P. Murray for discussions. This study was supported by two grants from the Rio Tinto WA Future Fund to the Western Australian Museum. Isotopic analysis was funded by a US National Science Foundation grant.

Author Contributions G.J.P. was responsible for site descriptions, faunal analysis, data synthesis and writing the paper, J.A.L. fieldwork coordination, L.K.A. stable isotope analysis, J.C.H. U-series dating, B.P. magnetic polarity assessment, M.N.H. lizard identifications, W.E.B. bird identifications, R.G.R., M.L.C. and L.J.A. optical dating, P.D.D. site interpretation, and N.M.W. small marsupial identifications.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to G.J.P. (gavin.prideaux@museum.wa.gov.au).

LETTERS

Retention of transcriptionally active cryptophyte nuclei by the ciliate *Myrionecta rubra*

Matthew D. Johnson[1]†, David Oldach[2], Charles F. Delwiche[3] & Diane K. Stoecker[1]

It is well documented that organelles can be retained and used by predatory organisms, but in most cases such sequestrations are limited to plastids of algal prey[1]. Furthermore, sequestrations of prey organelles are typically highly ephemeral[2] as a result of the inability of the organelle to remain functional in the absence of numerous nuclear-encoded genes involved in its regulation, division and function[3]. The marine photosynthetic ciliate *Myrionecta rubra* (Lohmann 1908) Jankowski 1976 (the same as *Mesodinium rubrum*)[4] is known to possess organelles of cryptophyte origin[5–9], which has led to debate concerning their status as permanent symbiotic or temporary sequestered fixtures[5–13]. Recently, *M. rubra* has been shown to steal plastids (that is, chloroplasts) from the cryptomonad, *Geminigera cryophila*, and prey nuclei were observed to accumulate after feeding[10]. Here we show that cryptophyte nuclei in *M. rubra* are retained for up to 30 days, are transcriptionally active and service plastids derived from multiple cryptophyte cells. Expression of a cryptophyte nuclear-encoded gene involved in plastid function declined in *M. rubra* as the sequestered nuclei disappeared from the population. Cytokinesis, plastid performance and their replication are dependent on recurrent stealing of cryptophyte nuclei. Karyoklepty (from Greek *karydi*, kernel; *kleftis*, thief) represents a previously unknown evolutionary strategy for acquiring biochemical potential.

We have previously shown that *M. rubra* ingests cryptophyte algae and retains organelles[10,12], leading to enhanced photosynthetic[10,11] and growth rates[10,11]. However, although cryptophyte nuclei have been observed in *M. rubra* after feeding[10,11], their viability or function has never been determined. The plastids of *M. rubra* are organized in numerous 'complexes'[5] that also contain cryptophyte mitochondria and cytoplasm, and they are packaged by a host membrane and two endoplasmic reticulum (ER) membranes[5,7–9]. When present, cryptophyte nuclei may be isolated in the cytosol (Fig. 1) or closely associated with one or more 'chloroplast–mitochondrial complexes' (Supplementary Fig. 1a) but are never found within them.

Geminigera cryophila cells are intact on ingestion (Fig. 1b), after which their membrane is compromised and organelles are sequestered. Newly ingested prey nuclei (3–4 μm) appear as they do in *G. cryophila* cells (Fig. 1a, inset); however, after 30 days 50% of cryptophyte nuclei in *M. rubra* increase in size to 7–10 μm (Fig. 1c, d, and Supplementary Fig. 2). When present, the double-membraned *G. cryophila* nucleus in *M. rubra* is surrounded by a single membrane, enclosing ribosome-rich cytoplasm and sometimes mitochondria, and closely associated with two ER membranes (Fig. 1c). Although the outer ER membrane surrounding nuclei and plastids in cryptophytes seems to be broken during the sequestration process, it is possible that the formation of ER connections between the two in *M. rubra* may facilitate protein secretion into organelle complexes. To verify that these nuclei were indeed from *G. cryophila* and to determine their fate, we applied a fluorescence *in situ* hybridization (FISH) probe for the cultured *G. cryophila* small-subunit (SSU) ribosomal RNA nuclear gene using techniques established previously[14]. The probe bound to RNA in the nucleolus of prey nuclei, to

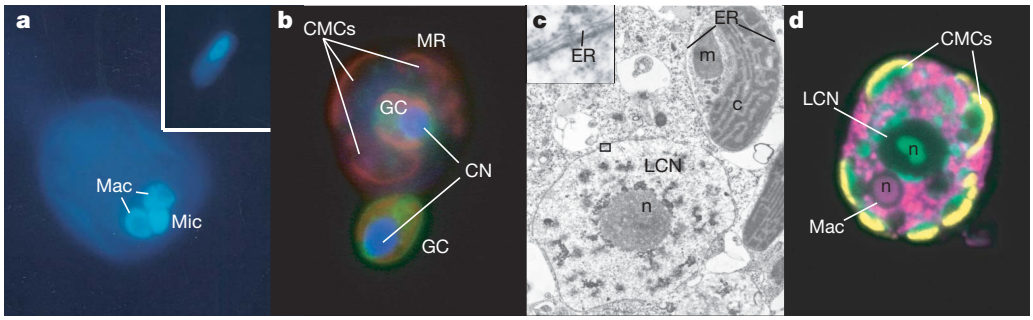


Figure 1 | Micrographs of *Myrionecta rubra* with *Geminigera cryophila* nuclei. **a**, Fluorescence micrograph of a DAPI-stained (blue) *Myrionecta rubra* cell without a cryptophyte nucleus and a free-living *Geminigera cryophila* cell (inset). Original magnification ×1,000. **b**, Layered fluorescence micrographs of a *M. rubra* cell with a newly ingested and free-living *G. cryophila* cell, hybridized with a FISH probe for *G. cryophila* SSU rRNA (green), stained with DAPI, and of endogenous plastid fluorescence (orange). Original magnification ×1,000. **c**, Transmission electron microscopy section of a *M. rubra* cell showing chloroplast–mitochondrial complexes and a cryptophyte nucleus, with detail of the surrounding membrane and ER (box and inset). Original magnification ×6,000. **d**, A layered three-channel (excitation at 488, 543 and 633 nm) confocal laser-scanning micrograph of a *M. rubra* cell dual-labelled with FISH probes for *G. cryophila* (green) and *M. rubra* (pink) SSU rRNA. Original magnification ×600. **c**, chloroplast; CMC, chloroplast–mitochondrial complex; CN, cryptophyte nucleus; GC, *G. cryophila* cell; LCN, large cryptophyte nucleus; m, mitochondrion; Mac, ciliate macronucleus; Mic, ciliate micronucleus; MR, *M. rubra* cell; n, nucleolus.

[1]University of Maryland Center for Environmental Science, Horn Point Laboratory, Cambridge, Maryland 21613, USA. [2]Institute of Human Virology, University of Maryland, School of Medicine, Baltimore, Maryland 21201, USA. [3]Cell Biology and Molecular Genetics, University of Maryland – College Park, College Park, Maryland 20742, USA. †Present address: Institute of Marine and Coastal Sciences, Rutgers University, 71 Dudley Road, New Brunswick, New Jersey 08901, USA.

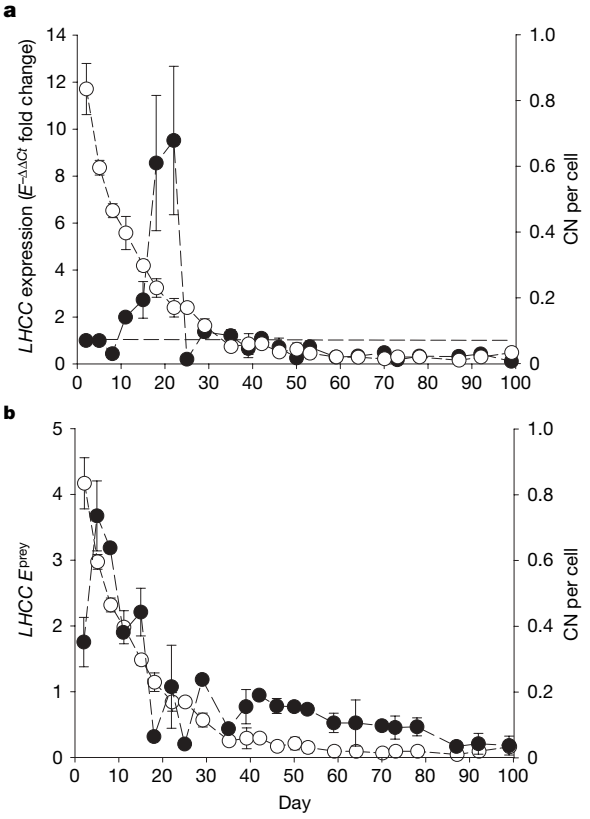


Figure 2 | Expression of the cryptophyte nuclear-encoded gene for the plastid-targeted protein *LHCC10* in *Myrionecta rubra*, and the presence of cryptophyte nuclei during starvation. a, Expression normalized to *M. rubra* β-tubulin gene expression ($E^{-\Delta\Delta Ct}$; filled circles)[15] shown with cryptophyte nuclei (CN) over time (open circles). The reference line across graph at 1 represents zero change from $t = 0$; below this line, expression has decreased from $t = 0$. **b**, Cryptophyte nuclei per cell (open circles) and *M. rubra LHCC* expression normalized to RNA standards of exponentially growing *G. cryophila* (E^{prey}), presented as equivalent *G. cryophila* cells per *M. rubra* cell (open circles). All results are means ± s.d.; $n = 2$.

cytoplasm enclosed within the ER surrounding the nucleus, to pockets of sequestered cytoplasm within the ciliate, and to the chloroplast–mitochondrial complexes (Fig. 1d). A FISH probe for the *M. rubra* nuclear SSU rRNA gene[14] was used in conjunction with the *G. cryophila* probe, and together they illustrate the mosaic nature of cryptophyte and endogenous cytoplasm in *M. rubra* (Fig. 1d; see also Supplementary Fig. 1a).

Gene expression of the sequestered nuclei in *M. rubra* includes plastid-targeted protein genes. Overexpression of the light-harvesting chloroplast complex protein gene (*LHCC10*)[15] was observed during the first 20–30 days after feeding (cultures fed 7 days before $t = 0$). When normalized to a *M. rubra* housekeeping gene[16], expression of *LHCC10* was enhanced up to tenfold in the first 20 days relative to $t = 0$ (Fig. 2a), and peaked when most sequestered nuclei had become unusually large (Supplementary Fig. 2a). Expression normalized to cryptophyte-only RNA standards was threefold to fourfold that of a cryptophyte cell during the same time and closely followed changes in cryptophyte nuclei per cell (Fig. 2b). The cryptophyte genes for the D1 protein (*psbA*, plastid) and the regulatory protein for ribulose bisphosphate carboxylase/oxygenase (RUBISCO) (*CbbX*, nucleomorph[17]) both revealed maximum expression during the first 2 weeks after the ingestion of cryptophyte prey (Supplementary Fig. 3). However, expression of the *CbbX* gene declined to relatively low expression levels after 3 weeks, whereas the plastid *psbA* gene maintained higher expression levels.

There was no evidence of net division of cryptophyte nuclei, although several sequestered nuclei were observed in karyokinesis (Supplementary Fig. 1b). Because the net retention time of plastids is greater than that for prey nuclei, an average *M. rubra* cell may have eight cryptophyte plastids per single prey nucleus. Thus, individual cryptophyte nuclei in *M. rubra* cells apparently service plastids originating from several *G. cryophila* cells. The decline in number of cryptophyte nuclei per cell followed an exponential decay curve ($r^2 = 0.983$) with a half-life of 9.53 days and a maximum retention time of 30 days (Supplementary Fig. 4). By day 35, cytokinesis of *M. rubra* decreased to half of levels at $t = 0$, approaching zero by the end of the experiment (Table 1). Loss of regulatory control over sequestered plastids lagged the loss of prey nuclei, with significant declines in plastid number per cell and photosynthetic quantum efficiency (F_v/F_m, where F_v is variable fluorescence and F_m is maximum fluorescence) after day 74 (Table 1).

These data show that sequestered cryptophyte nuclei are transcriptionally active in *M. rubra* and that cryptophyte organelles function 'symbiotically' during the period of nuclear retention. However, the destruction of cryptophyte prey cells precludes interpreting the relationship as symbiosis, and the process is best characterized as predation with farming of the prey organelles. Plastids in *M. rubra* do not seem to be permanently integrated cellular features. Prey nuclei are stolen and replaced by nearly constant feeding on cryptophyte algae; the nuclei seem to maintain the more stable plastids and mitochondria. Such a strategy may minimize predator investment in maintenance of the symbiont. Loss of prey nuclei results in the inability to divide plastids, leading to declines in organelle concentrations and biochemical potential (Table 1). Although the least stable aspect of this survival strategy seems to be the prey nucleus, the acquisition of new prey nuclei through feeding is potentially about 1 per day under natural conditions. The potential feeding rate is of the same order as cytokinesis rates and is much shorter than the observed half-life (about 10 days) of cryptophyte nuclei in the ciliate. Thus, in feeding *M. rubra* populations, retained prey nuclei could be present nearly continuously.

Whereas no organism has ever been described to sequester nuclei of another, red algal adelphoparasites are known to deliver a nucleus into their host cell cytoplasm, where it undergoes DNA synthesis and karyokinesis within the host's cytoplasm[18]. However, the relationship between *M. rubra* and *G. cryophila* is strikingly different from that between red algae and their adelphoparasites because of the distant phylogenetic relationships between host and prey. Foreign nuclei have also been observed transiently in some plastid-retaining dinoflagellates[19–21], but their role in the cells, if any, is unknown. Present-day observations of organelle retention, symbiosis and parasitism offer dynamic pictures of interspecies organellar and genomic interactions, and help in understanding the complex evolutionary history of eukaryotic cells. Although we cannot say whether nuclear retention

Table 1 | Physiological parameters for *Myrionecta rubra* during starvation

Period	Days	Growth per day	Plastid division per day	Nucleomorph genome per cell	Photosynthetic quantum efficiency
1	0–18	0.073 ± 0.019	0.075 ± 0.026	8.6 ± 1.4	0.61 ± 0.018
2	19–35	0.071 ± 0.016	0.061 ± 0.005	9.3 ± 1.2	0.61 ± 0.021
3	36–53	0.032 ± 0.002*	0.018 ± 0.009*	7.8 ± 1.2	0.60 ± 0.023
4	54–73	0.035 ± 0.001*	0.021 ± 0.001*	6.0 ± 1.0*	0.59 ± 0.038
5	74–99	0.015 ± 0.010*	0.009 ± 0.022*	3.9 ± 0.7*	0.49 ± 0.046*

All results are means ± s.d. ($n = 2$) over each period. Periods are defined by transfer to new medium. Nucleomorph genome per cell approximates to plastids per cell. Asterisk indicates a significant difference ($P < 0.05$) from period 1 values (one-tailed analysis of variance) with Tukey's Studentized range test.

is an evolutionary step towards the permanent establishment of a genetically integrated plastid, it offers a striking example of cellular chimaerism and has proved to be a successful ecological phenomenon. *M. rubra* is a highly successful and widespread member of the plankton, capable of forming dense red tides in coastal and upwelling waters[5,6]. Karyoklepty is a unique attribute, allowing temporary access to the genetic information and biochemical potential of another species.

METHODS

Experiment preparation. *Myrionecta rubra* (CCMP 2563) and *Geminigera cryophila* (CCMP 2564) cultures were maintained as described previously[11]. *G. cryophila* were added one week before $t = 0$ to semi-continuous batch cultures ($n = 2$) at a 4:1 ratio for the gene expression experiment. Sampling, measurement of growth rates, and nuclei counts by staining with 4′,6-diamidino-2-phenylindole (DAPI) were conducted with techniques described previously[10,11]. Transmission electron microscopy was performed, and F_v/F_m was measured, as described previously[12].

FISH. A FISH oligonucleotide probe, *TANU2*, labelled with fluorescein isothiocyanate (FITC) (Supplementary Table 1), for the *G. cryophila* SSU rRNA gene was designed by eye from DNA alignments with MacClade 4.05 (ref. 22). FISH probes for the *M. rubra* SSU rRNA gene (*MYR2*), labelled with 5-N,N′-diethyltetramethylindodicarbocyanine (Cy5)[14], positive control (uniC-FITC), and negative (anti-sense) control (uniR-FITC) probes[23] were also used. All techniques used were as described previously[14].

DNA extraction, polymerase chain reaction (PCR) and sequencing. DNA extraction, PCR amplification and gene sequencing for *M. rubra* genes were conducted with methods outlined previously[14]. The β-tubulin gene (*β-tub*) was isolated from *M. rubra* by using universal primers[24]. The light-harvesting complex protein (*LHCC10*) was isolated with primers (Supplementary Table 1) designed by eye from gene alignments in MacClade 4.05 (ref. 22).

RNA isolation, quantitative PCR and RT–PCR. RNA was isolated with the RNeasy Plant Mini Kit (Qiagen) and treated with DNase before use for RT–PCR (see Supplementary Information for details). Quantitative PCR was used to quantify nucleomorph (Nm) genome content in *M. rubra*, as a proxy for plastid number, with methods described previously[12]. Creation of complementary DNA and quantitative PCR for measurements of gene expression were conducted with TaqMan assays and a Cepheid Smart Cycler (see Supplementary Information for details).

Received 15 August; accepted 30 November 2006.

1. Blackbourn, D. J., Taylor, F. J. R. & Blackbourn, J. Foreign organelle retention by ciliates. *J. Protozool.* **20,** 286–288 (1973).
2. Stoecker, D. K. & Silver, M. W. Replacement of aging chloroplasts in *Strombidium capitatum* (Ciliophora: Oligotrichida). *Mar. Biol.* **107,** 491–502 (1990).
3. Martin, W. *et al.* Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* **393,** 162–165 (1998).
4. Lynn, D. H. & Small, E. B. In *Illustrated Guide to the Protozoa* (eds Lee, J. J., Leedale, G. F. & Bradbury, P.) 477–478 (Society of Protozoologists, Lawrence, Kansas, 2000).
5. Taylor, F. J. R., Blackbourn, D. J. & Blackbourn, J. The red-water ciliate *Mesodinium rubrum* and its 'incomplete symbionts': a review including new ultrastructural observations. *J. Fish. Res. Bd Can.* **28,** 391–407 (1971).
6. Lindholm, T. *Mesodinium rubrum*—a unique photosynthetic ciliate. *Adv. Aquat. Microbiol.* **3,** 1–48 (1985).
7. Taylor, F. J. R., Blackbourn, D. J. & Blackbourn, J. Ultrastructure of the chloroplasts and associated structures within the marine ciliate *Mesodinium rubrum* (Lohmann). *Nature* **224,** 819–821 (1969).
8. Hibberd, D. J. Observations on the ultrastructure of the cryptomonad endosymbiont of the red water ciliate *Mesodinium rubrum*. *J. Mar. Biol. Assoc. UK* **57,** 45–61 (1977).
9. Oakley, B. R. & Taylor, F. J. R. Evidence for a new type of endosymbiotic organization in a population of the ciliate *Mesodinium rubrum* from British Columbia. *Biosystems* **10,** 361–369 (1978).
10. Gustafson, D. E., Stoecker, D. K., Johnson, M. D., Van Heukelem, W. F. & Sneider, K. Cryptophyte algae are robbed of their organelles by the marine ciliate *Mesodinium rubrum*. *Nature* **405,** 1049–1052 (2000).
11. Johnson, M. D. & Stoecker, D. K. The role of feeding in growth and the photophysiology of *Myrionecta rubra*. *Aquat. Microb. Ecol.* **39,** 303–312 (2005).
12. Johnson, M. D., Tengs, T., Oldach, D. & Stoecker, D. K. Sequestration, performance and functional control of cryptophyte plastids in the ciliate *Myrionecta rubra* (Ciliophora). *J. Phycol.* **42,** 1235–1246 (2006).
13. Hansen, P. J. & Fenchel, T. The bloom-forming ciliate *Mesodinium rubrum* harbours a single permanent endosymbiont. *Mar. Biol. Res.* **2,** 169–177 (2006).
14. Johnson, M. D., Tengs, T., Oldach, D. W., Delwiche, C. F. & Stoecker, D. K. Highly divergent SSU rRNA genes found in the marine ciliates *Myrionecta rubra* and *Mesodinium pulex*. *Protist* **155,** 347–359 (2004).
15. Deane, J. A. *et al.* Evidence for nucleomorph to host nucleus gene transfer: light-harvesting complex proteins from cryptomonads and chlorarachniophytes. *Protist* **151,** 239–252 (2000).
16. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *Methods* **25,** 402–408 (2001).
17. Gillott, M. A. & Gibbs, S. P. The cryptophyte nucleomorph: its ultrastructure and evolutionary significance. *J. Phycol.* **16,** 558–568 (1980).
18. Goff, L. J. & Coleman, A. W. Fate of parasite and host organelle DNA during cellular transformation of red algae by their parasites. *Plant Cell* **7,** 1899–1911 (1995).
19. Wilcox, L. W. & Wedemayer, G. J. *Gymnodinium acidotum* Nygaard (Pyrrophyta), a dinoflagellate with an endosymbiotic cryptomonad. *J. Phycol.* **20,** 236–242 (1984).
20. Fields, S. D. & Rhodes, R. G. Ingestion and retention of *Chroomonas* spp. (Cryptophyceae) by *Gymnodinium acidotum* (Dinophyceae). *J. Phycol.* **27,** 525–529 (1991).
21. Gast, R. J., Moran, D. M., Dennett, M. R. & Caron, D. A. Kleptoplastidy in an Antarctic dinoflagellate: caught in evolutionary transition? *Environ. Microb.* advance online publication, doi:10.1111/j.1462-2920.2006.01109.x (7 August, 2006).
22. Maddison, W. P. & Maddison, D. R. *MacClade—Analysis of Phylogeny and Character Evolution* (Sinauer, Sunderland, Massachusetts, 1992).
23. Miller, P. E. & Scholin, C. A. Identification and enumeration of cultured and wild *Pseudo-nitzschia* (Bacillariophyceae) using species-specific LSU rRNA-targeted fluorescent probes and filter-based whole cell hybridization. *J. Phycol.* **34,** 371–382 (1998).
24. Saldarriaga, J. F., McEwan, M. L., Fast, N. M., Taylor, F. J. R. & Keeling, P. J. Multiple protein phylogenies show that *Oxyrrhis marina* and *Perkinsus marinus* are early branches of the dinoflagellate lineage. *Int. J. Syst. Evol. Microbiol.* **53,** 355–365 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank T. Kana, D.W. Coats, K. Bidle and P. Falkowski for comments on this manuscript; D. Gustafson, S. Heyward and H. Bowers for advice and/or technical assistance; T. Kana for use of his PAM fluorimeter; and C. Scholin for assistance with FISH protocols. This project was funded by a National Science Foundation grant (to D.K.S.).

Author Contributions M.D.J. and D.K.S. conceived of the project. M.D.J. conducted all laboratory experiments and data analysis for the paper. D.K.S., D.O. and C.F.D. provided methodological expertise and contributed to the interpretation of data. M.D.J. wrote most of the paper, with contributions and advice from D.K.S., D.O. and C.F.D.

Author Information The sequences for the β-tubulin gene, *CbbX*, *LHCC10* and *psbA* are deposited in GenBank under accession numbers EF151014, EF151015, EF151016 and EF151017. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.D.J. (johnson@marine.rutgers.edu).

LETTERS

Fish can infer social rank by observation alone

Logan Grosenick[1,2]†, Tricia S. Clement[1]† & Russell D. Fernald[1]

Transitive inference (TI) involves using known relationships to deduce unknown ones (for example, using $A > B$ and $B > C$ to infer $A > C$), and is thus essential to logical reasoning. First described as a developmental milestone in children[1], TI has since been reported in nonhuman primates[2–4], rats[5,6] and birds[7–10]. Still, how animals acquire and represent transitive relationships and why such abilities might have evolved remain open problems. Here we show that male fish (*Astatotilapia burtoni*) can successfully make inferences on a hierarchy implied by pairwise fights between rival males. These fish learned the implied hierarchy vicariously (as 'bystanders'), by watching fights between rivals arranged around them in separate tank units. Our findings show that fish use TI when trained on socially relevant stimuli, and that they can make such inferences by using indirect information alone. Further, these bystanders seem to have both spatial and featural representations related to rival abilities, which they can use to make correct inferences depending on what kind of information is available to them. Beyond extending TI to fish and experimentally demonstrating indirect TI learning in animals, these results indicate that a universal mechanism underlying TI is unlikely. Rather, animals probably use multiple domain-specific representations adapted to different social and ecological pressures that they encounter during the course of their natural lives.

Territorial *A. burtoni* males engage in regular aggressive bouts that determine their access to territory and resources. Males that repeatedly lose fights are unable to hold territories and consequently descend in social status. These males display a different morphology from successful territorial males, losing their bright coloration and becoming reproductively dormant[11–13]. Success in aggressive bouts is therefore crucial to male reproductive fitness, and the ability to infer the relative strength of rivals before engaging them in potentially costly fights should be highly adaptive[13,14].

We tested whether bystander males, after observing pairwise fights between rivals, could synthesize this information to make inferences on an overall hierarchy implied by the fights. To do this we used an adaptation[2] of a classic five-element model developed to test children's ability to make TIs on asymmetric relationships[15]. Each bystander fish saw staged fights between five size-matched males (A to E, where each letter stands for a different rival male). During training, the following staged fights were presented to the bystander: A+B−, B+C−, C+D− and D+E−, where a plus denotes a 'win', and a minus a 'loss' (for example, in the pair A+B−, rival A dominates rival B) (see Methods). These fights, taken together, imply the dominance hierarchy A > B > C > D > E.

Each bystander was trained in the central unit of a square tank divided into several visually and chemically isolated units (Fig. 1). Fights were staged by moving one rival into another rival's unit, and then removing an opaque barrier, making the fight visible to the bystander (the fight A+B− is shown diagrammatically in Fig. 1). Because *A. burtoni* individuals vigorously defend their territory against intruding rivals, moving one male into a unit defended by another male always resulted in the intruder losing (see Methods). Thus, we could train each bystander on an artificial dominance hierarchy by using animals whose relative status we controlled. This ensured that there were no consistent differences in male abilities or physical characteristics—a potential confounding factor in naturally formed dominance hierarchies[14,15].

Bystander males ($n = 8$) were trained for 11 days on pairwise fights implying A > B > C > D > E (see Methods). Then, to assess whether they could make inferences on the implied hierarchy, we tested their preference between rivals that they had never seen together, giving them a choice between rivals A and E ('AE') and a choice between B and D ('BD'). In the staged fights presented to the bystander, A had always won (against B) and E had always lost (to D). A comparison between these hierarchy end points, known in the literature as 'end anchors', was therefore included as a baseline assay of bystander preference (because end anchors are known to generate a large, unambiguous response in TI tasks[16]). The BD choice assayed for an inference on the hierarchy, because the B and D rivals won and lost the same number of fights during training, and therefore, after appropriate controls for physical and spatial differences, differed only in their relative position in the implied hierarchy (see Methods).

During bystander preference testing, each bystander was placed between the members of one of the novel pairs (either AE or BD) and, after a controlled viewing period, was allowed to swim freely

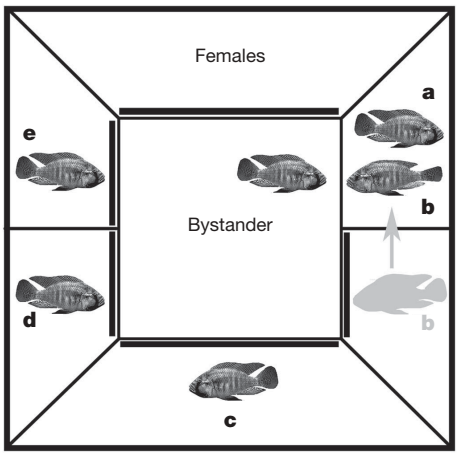


Figure 1 | Tank arrangement and bystander training. Five rival males (A, B, C, D and E) were arranged in visually, chemically and physically isolated compartments around the central bystander unit. To train a bystander on a particular fight, the male scheduled to be the 'loser' was removed from its unit and placed in the territory of the scheduled 'winner'. The opaque barrier separating the bystander from the rivals was then removed to allow the bystander to view the fight. The fight A+B− (A 'wins', B 'loses') is shown here in diagrammatic form.

[1]Department of Biological Sciences, Stanford University, Stanford, California, 94305, USA. [2]Center for the Study of Language and Information, Stanford University, Stanford, California, 94305, USA. †Present addresses: Center for the Study of Language and Information, Ventura Hall, 200 Panama Street, Stanford, California 94305, USA (L.G.); eBay, Inc., 2145 Hamilton Avenue, San Jose, California 95125, USA (T.S.C.).

between them. We recorded both the simple binary first-choice result (which rival the bystander approached first) and the overall time spent adjacent to each rival (see Methods). Previous experiments in *A. burtoni*[17] and other fish[14] have shown that time spent in tank quadrants adjacent to a particular male indicates bystander 'preference', and that bystanders spend more time near the rival they perceive to be weaker[14]. This is consistent with previous studies of preference in dominance models and, along with our own unpublished findings on *A. burtoni*, motivated our a priori hypothesis that bystanders would prefer the lower-ranking rival in this experiment. Each bystander ($n = 8$) was tested for TI in both the tank where it had seen the staged fights ('familiar context'), and in a tank it had never been in before ('novel context'; see Methods). We tested half the rivals in the familiar context first, and the other half in the novel context first. We used the same AE and BD rivals in both contexts (see Methods).

We used the first-choice results in a general assay for TI, using one-tailed hypothesis testing in the direction of our a priori hypothesis and a significance level of $\alpha = 0.05$. For the AE choice in the familiar context, E was chosen eight out of eight times ($P = 0.004$, binomial), showing that bystanders prefer losing rivals to winning ones here, as expected. In the novel context, however, E was chosen only six out of eight times ($P = 0.14$, binomial). On the critical BD choice, D was chosen seven out of eight times in the familiar context ($P = 0.035$, binomial), and eight out of eight times in the novel context ($P = 0.004$, binomial). Therefore, bystanders chose D as though it were the losing fish, even though (unlike A and E) the B and D rivals differed only in their relative position in the implied hierarchy. Still, this initial evidence on a binary variable was non-significant in one AE context and tells us little about behaviour during preference. We therefore turned to the continuous 'time spent' data to look for subtler effects.

As our sample was too small for appropriate validation of the usual assumptions of parametric significance testing, we used a nonparametric permutation *t*-test (50,000 replicates) that did not make unverifiable parametric assumptions about the underlying distribution of the data (see Methods). Testing was one-tailed in the direction of our a priori hypothesis, with a significance level of $\alpha = 0.01$. The time spent (T_s) adjacent to each rival was summarized in one score as the difference between time spent near the lower-ranking rival minus the time spent near the higher-ranking rival. Under the null hypothesis that bystanders show no preference for the lower-ranking rival, the expectation of T_s would be less than or equal to zero.

In the A versus E preference test ($n = 8$), T_s was significantly positive in both the familiar context ($P = 0.01$) and the novel context ($P = 0.003$), showing that the bystanders significantly preferred E as expected. In the critical B versus D preference test, T_s was again significantly positive in both the familiar context ($P = 0.003$) and in the novel context ($P = 0.008$), showing a significant preference for the D rival in both contexts (see Fig. 2).

We wondered whether the degree to which one rival dominated over or submitted to other rivals during bystander training had an impact on bystander inference behaviour. Specifically, if the C rival's dominant and submissive behaviour were ambiguous, there might be less social information available to the bystander about the transitive relationship B > D than if the outcomes of the fights were clear (that is, if C very obviously submitted to B and dominated D).

To examine such effects we calculated a 'dominance score' (DS) measuring each rival's dominance/submission behaviour in each fight. Focal observations of five highly correlated behavioural variables known to index dominant and submissive behaviour (chase, flee, bite, threaten, and eyebar activation)[13] collected from both rivals during each of the 336 fights were subjected to a standard principal-components analysis (PCA). Data projected onto the first principle component then provided a single DS for each rival male in each fight. DS values were more positive for increasing dominant behaviour, more negative for increasing submissive behaviour, and near zero for ambiguous behaviour (see Methods).

To explore whether rival dominant/submissive activity predicted bystander performance on the BD discrimination, we regressed T_s for the BD preference choice on a combination of C dominant/submissive behaviours as seen by each bystander (combined rival behaviour). Combined rival behaviour was calculated for each bystander from the normalized absolute values of the summed DS values (it therefore indexed overall activity, independently of whether that activity was dominant or submissive; see Methods). Given the number of data points ($n = 8$) and multiple tests, it was decided beforehand that all P values would be Bonferroni-corrected to account for familywise error and then tested at the $\alpha = 0.05$ level (see Methods). We then ran regressions separately for the familiar and novel contexts. Interestingly, we found that while C rival activity during training failed to predict bystander discrimination in the familiar context ($t = -1.271$, $P = 1$ (corrected), $R = 0.461$), it did so to a significant extent in the novel context ($t = 3.896$, $P = 0.048$ (corrected), $R = 0.847$), positively predicting time spent away from B and near D (Fig. 3). Furthermore, a comparison of the regression slopes using repeated-measures analysis of covariance indicated that the two regression slopes were significantly different at the $\alpha = 0.05$ level ($F_{2,12} = 6.50$, $P = 0.012$). Corresponding regressions on B and D rival activity were not significant for either context.

Previous research has shown that fish use redundant coding of spatial and featural information to represent their surroundings[18]. We suggest that such redundant representation is behind the pronounced difference in how social information learned by bystanders during training affects their performance in different contexts during testing. In particular, it is clear that when information about the training context was available to them (as in the familiar context), all bystanders were able to make the correct discrimination. However, given only rival featural information (as in the novel context), bystanders showed a linearly graded performance based on how clear fight outcomes involving the transitive element C had been. This strongly suggests that different representations of rival abilities related to learned spatial and featural cues may be used by bystander males, depending on what contextual information is available to them at the time of choice.

The mechanism underlying TI behaviour in animals has been widely debated[2,4–10,19–25], and often acts as a proxy for a larger debate between proponents of associative versus cognitive accounts of animal learning and behaviour[26]. In associative models, TI behaviour is a result of direct reinforcement history. Such models cannot explain the current findings without modification, because vicarious TI involves no direct reinforcement of any kind. Further, the linear relationship observed between transitive C element activity and T_s

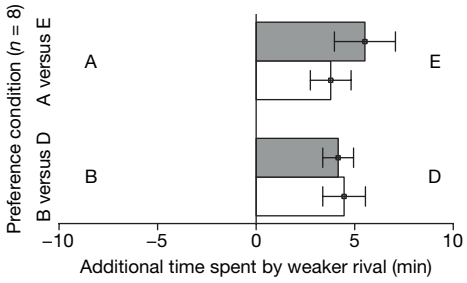


Figure 2 | Time spent near rival males during preference testing. Bar plots of the difference in time spent near the higher-ranking versus the lower-ranking rival (T_s) in the AE (top) and BD (bottom) preference tasks by context. Filled bars, familiar context; open bars, novel context. One-sample permutation *t*-tests (50,000 replicates) showed that in all cases bystanders spent significantly more time near the lower-ranking rival, namely D or E (AE familiar context, $P = 0.01$; AE novel context, $P = 0.003$; BD familiar context, $P = 0.003$; BD novel context, $P = 0.008$). Error bars indicate s.e.m.

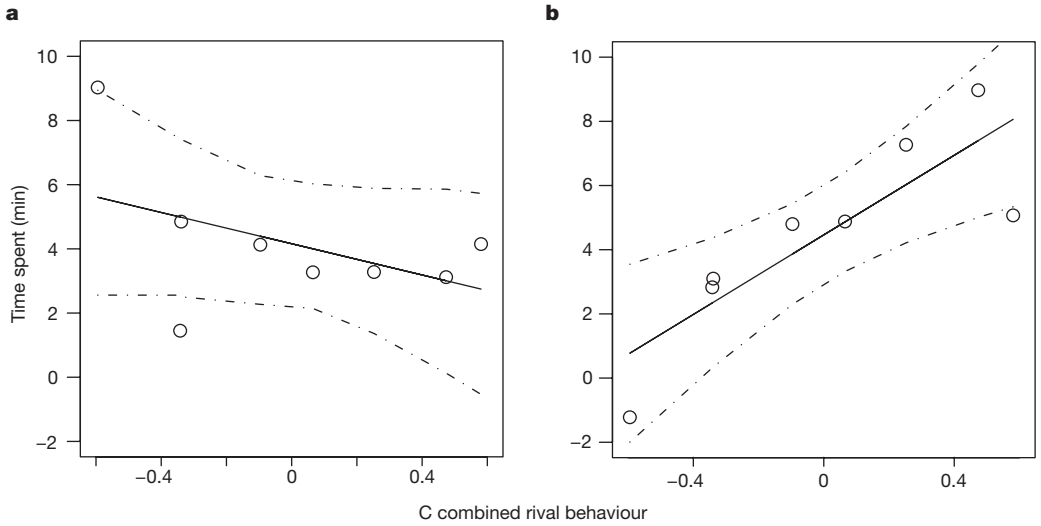


Figure 3 | Time spent during BD preference task as predicted by C combined rival behaviour. **a**, Familiar context; **b**, novel context. T_s, the time spent in the quadrant nearest B minus the time spent in the quadrant nearest D, was regressed on C combined rival behaviour, a measure of overall C rival behaviour during fights. The plots show that whereas C combined rival behaviour failed to predict T_s in the familiar context ($P = 1$; $R = -0.461$), it did so to a significantly positive extent in the novel context ($P = 0.048$; $R = 0.847$). Both P values have been Bonferroni corrected for multiple comparisons. Dotted lines represent 95% confidence intervals for the line.

in the novel context (when no such relationships were seen for B and D activity) opposes associative 'value transfer' models, in which values associated with B and D would be determined by the behaviour of neighbouring end anchors A and E[8,20]. We conclude that, at least in this context, a more complex representation is at work.

Last, this instance of the same animal using different representations for TI given different contextual information suggests that considering a single underlying mechanism for TI may not be sufficient. Instead, it seems likely that animals have developed domain-specific representations adapted to the particular social and ecological pressures that they encounter in their natural lives. It has already been established that animals use TI in social settings[27], and for *A. burtoni* separate spatial and featural representations underlying TI would be quite consistent with males' different social and ecological needs. In their native habitat in the estuaries and temporary shore-pools of Lake Tanganyika, *A. burtoni* individuals find their established territories regularly disturbed by wind, predation, the movements of hippopotamuses, or the natural changing of the shore-pools[11,12]. With old territory boundaries destroyed, rival featural representations independent of the altered spatial context would be invaluable—especially because such times of instability have been shown to allow rapid ascent (or descent) in dominance status and therefore in corresponding physiology and reproductive success[28,29]. Meanwhile, in a stable setting, an additional territorial representation incorporating more than just the degree of rival wins and losses seems to allow more robust TI performance given ambiguous social information—apparently exploiting the spatial acuity of this highly territorial animal.

METHODS

Training tank arrangement, and controls. Rivals ($n = 10$) and bystanders ($n = 8$) were matched for size and weight by tank, and had never met before. Once selected, bystanders and rivals were housed in a square purpose-built tank ($76.2 \times 76.2 \times 25.4\ cm^3$; Fig. 1). In each of four runs, two bystander fish were placed in the divided centre compartment of the tank, and the five rival fish (A, B, C, D and E) were arranged in units surrounding this central observation unit (Fig. 1). Two pairs of bystanders were trained on $n = 5$ rivals each ($n = 10$ total), and then two more pairs were trained on the same rivals with reversed hierarchy order (see below). Further details are provided in Supplementary Information.

To control for stable physiological differences between fish chosen to be A–E rivals, rival positions were exchanged such that fish serving as the A rival for half of the bystanders served as the E rival for the other half. The same was true for the B and D rivals. To control for possible spatial effects resulting from consistent location of A next to B and D next to E in the arrangement of rivals around the central bystanders, we also exchanged the spatial locations of the A and E rivals (see Supplementary Information). This spatial exchange was counterbalanced across the position exchange. Together these controls left only fights and transient rival behaviour to influence the bystanders' AE and BD preference, thus controlling for any consistent differences between rival males in physiology, behaviour, or location.

Bystander training on fights between rival males. Bystanders were trained for 7 min per fight and saw two fights a day, one in the morning 2 h after feeding and one 4 h later, for 11 consecutive days (except on days 5, 10 and 11, when bystanders viewed all four possible pairwise fights—see Supplementary Information for complete training details).

At the start of a training trial the rival fish designated to lose was removed from its territory, stressed twice by suspension out of water for 30 s (separated by 1 min) and then placed in the territory of the rival designated to win. The stress, combined with invader status, guaranteed an unambiguous intruder loss. At 1 min after the intruding rival's introduction into the winner's territory, the opaque barrier separating the observer from the competing rivals was removed, and behavioural data were recorded for 7 min. After 7 min the barrier was replaced and the losing rival was returned to its territory, ending the fight.

Bystander preference between novel pairs. Preference during visual presentation of AE and BD pairings was assessed with the use of an approach/avoidance task conducted both in the 'familiar context' (training tank) (bystander choice area: length 45.72 cm, depth 22.86 cm, height 27.94 cm) and in a 'novel context' (unfamiliar tank) specifically designed for testing the approach/avoidance preference in fish[17] (bystander choice area: length 74 cm, depth 37 cm, height 28 cm) (Supplementary Fig. 1). Testing order was counterbalanced, and bystanders were tested on the same pair of rivals in both contexts. Possible effects of test tank differences are discussed in Supplementary Information. During bystander training on pairwise fights suggesting the A > B > C > D > E hierarchy, the order in which bystanders saw B and D lose was counterbalanced to avoid primacy or recency effects; the same was true for A and E (see Supplementary Information).

Each assessment included the following: first, a baseline period, during which the bystander was allowed to swim freely about the tank, but rivals were not visible; second, a forced viewing period during which the bystanders movement was limited to a $4 \times 4 \times 6\ ft^3$ (1 ft = 0.3048 m) clear plastic box, placed equidistant from visible rival males; and third, a preference period, during which the bystander was allowed to swim freely between the visible rivals. Each period lasted 10 min and scoring was double-blind. T_s was measured between the two farthest quadrants during the preference period, because fish spent nearly all their time in these quadrants when rivals were visible. Further information on preference testing and related controls is provided in Supplementary Information.

Behavioural observations and analysis. *A. burtoni* males have stereotypic patterns of behaviour that reflect their dominance status. In particular, five behavioural variables: chase, bite, flee, lateral threat, and eyebar activation are good indicators of male dominance status[11–13]. Qualitative 'winning' and 'losing', as discussed above, were thus defined behaviourally. Losers consistently flee and downregulate their eyebar, whereas winners chase, bite, threaten, and upregulate their own eyebar activation. Overall behavioural differences were quantified with PCA, yielding a DS for each fish in each fight. In all 672 cases, the designated 'winner' had a positive DS and the designated 'loser' a negative DS. More information on the PCA is provided in Supplementary Information. Permutation *t*-tests and univariate regressions were standard and are discussed in detail in Supplementary Information. *R* was used for plotting and statistical analysis[30].

Received 7 August; accepted 5 December 2006.

1. Piaget, J. *Judgement and Reasoning in the Child* (Kegan, Paul, Trench & Trubner, London, 1928).
2. McGonigle, B. O. & Chalmers, M. Are monkeys logical? *Nature* **267,** 694–696 (1977).
3. Rapp, P. R., Kansky, M. T. & Eichenbaum, H. Learning and memory for hierarchical relationships in the monkey: Effects of aging. *Behav. Neurosci.* **110,** 887–897 (1996).
4. Gillian, D. J. Reasoning in the chimpanzee: II. Transitive inference. *J. Exp. Psychol. Anim. Behav. Process.* **7,** 87–108 (1981).
5. Davis, H. Transitive inference in rats (*Rattus norvegicus*). *J. Comp. Psychol.* **106,** 342–349 (1992).
6. Roberts, W. A. & Phelps, M. T. Transitive inference in rats—a test of the spatial coding hypothesis. *Psychol. Sci.* **5,** 368–374 (1994).
7. von Fersen, L., Wynne, C. D., Delius, J. D. & Staddon, J. E. Transitive inference formation in pigeons. *J. Exp. Psychol. Anim. Behav. Process.* **17,** 334–341 (1991).
8. Steirn, J. N., Weaver, J. E. & Zentall, T. R. Transitive inference in pigeons: Simplified procedures and a test of value transfer theory. *Anim. Learn. Behav.* **23,** 76–82 (1995).
9. Bond, A. B., Kamil, A. C. & Balda, R. P. Social complexity and transitive inference in corvids. *Anim. Behav.* **65,** 479–487 (2003).
10. Lazareva, O. F. *et al.* Transitive responding in hooded crows requires linearly ordered stimuli. *J. Exp. Anal. Behav.* **82,** 1–19 (2004).
11. Fernald, R. D. & Hirata, N. R. Field study of *Haplochromis burtoni*: Habitats and cohabitants. *Environ. Biol. Fishes* **2,** 299–308 (1977).
12. Fernald, R. D. & Hirata, N. R. Field study of *Haplochromis burtoni*: Quantitative behavioral observations. *Anim. Behav.* **25,** 964–975 (1977).
13. Fernald, R. D. Quantitative behavioural observations of *Haplochromis burtoni* under semi-natural conditions. *Anim. Behav.* **25,** 643–653 (1977).
14. Oliveira, R. F., McGregor, P. & Latruffe, C. Know thine enemy: fighting fish gather information from observing conspecific interactions. *Proc. R. Soc. Lond. B* **265,** 1045–1049 (1998).
15. Chase, I. D. Individual differences versus social dynamics in the formation of animal dominance hierarchies. *Proc. Natl Acad. Sci. USA* **99,** 5744–5749 (2002).
16. Bryant, P. E. & Trabasso, T. Transitive inferences and memory in young children. *Nature* **232,** 456–458 (1971).
17. Clement, T. S., Grens, K. E. & Fernald, R. D. Female affiliative preference depends on reproductive state in the African cichlid fish, *A. burtoni*. *Behav. Ecol.* **16,** 83–88 (2005).
18. Vargas, J. P., Lopez, J. C., Salas, C. & Thinus-Blanc, C. Encoding of geometric and featural spatial information by goldfish (*Carassius auratus*). *J. Comp. Psychol.* **118,** 206–216 (2004).
19. De Lillo, C., Floreano, D. & Antinucci, F. Transitive choices by a simple, fully connected, backpropagation neural network: implications for the comparative study of transitive inference. *Anim. Cogn.* **4,** 61–68 (2001).
20. von Fersen, L., Wynne, C. D. L., Delius, J. & Staddon, J. E. R. Transitive inference formation in pigeons. *J. Exp. Psychol. Anim. Behav. Process.* **17,** 334–341 (1991).
21. Wynne, C. D. in *Models of Action: Mechanisms for Adaptive Behavior* (eds Wynne, C. D. & Staddon J. E.) 269–307 (Lawrence Erlbaum, Mahwah, NJ, 1998).
22. McGonigle, B. O. & Chalmers, M. in *Reasoning and Discourse Processes* (eds Myers, T., Brown, K., & McGonigle, B. O.) 141–164 (Academic, London, 1986).
23. McGonigle, B. O. & Chalmers, M. Monkeys are rational! *Q. J. Exp. Psychol.* **45B,** 198–228 (1992).
24. Eichenbaum, H. & Dusek, J. A. The hippocampus and memory for orderly stimulus relations. *Proc. Natl Acad. Sci. USA* **94,** 7109–7114 (1997).
25. Zentall, T. R. The case for a cognitive approach to animal learning and behaviour. *Behav. Processes* **54,** 65–78 (2001).
26. Allen, C. in *Rational Animals?* (eds Hurley, S. & Nudds, M.) 175–185 (Oxford Univ. Press, Oxford, 2006).
27. Paz-Y-Mino, C. G., Bond, A. B., Kamil, A. C. & Balda, R. P. Pinyon jays use transitive inference to predict social dominance. *Nature* **430,** 778–781 (2004).
28. Hoffman, H. A., Benson, M. E. & Fernald, R. D. Social status regulates growth rate: Consequences for life-history strategies. *Proc. Natl Acad. Sci. USA* **96,** 14171–14176 (1999).
29. White, S. A., Nguyen, T. & Fernald, R. D. Social regulation of gene expression in male and females social status. *J. Exp. Biol.* **205,** 2567–2581 (2002).
30. R Development Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna) ⟨http://www.R-project.org⟩ (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Le for rescoring data for the reliability analysis, and T. Zentall, L. Harbott and P. Suppes for their comments. Supported by National Institutes of Health awards (R.D.F. and T.S.C.).

Author Contributions T.S.C. and L.G. were responsible for experimental design. L.G. was responsible for data collection and analysis. L.G. wrote the manuscript. T.S.C., R.D.F. and L.G discussed the experiment and edited the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to L.G. (logang@stanford.edu).

CORRIGENDUM

doi:10.1038/nature05642

Genetically modified *Plasmodium* parasites as a protective experimental malaria vaccine

A. K. Mueller, M. Labaied, S. Kappe & K. Matuschewski

Nature 433, 164–176 (2005)

It has been drawn to *Nature*'s attention that A.K. M., S. K. and K. M. were named as inventors on a patent application relevant to this work (patent number WO 2005/063991) in 2004, which should therefore have been formally declared as a competing financial interest. The authors' non-profit institutions filed for patent protection to promote the development and distribution of malaria vaccines to people in need worldwide, in accordance with a global access strategy[1].

1. Chen, I. Thinking big about global health. *Cell* 124, 661–663 (2006).

CORRIGENDUM

doi:10.1038/nature05640

Deletion of active ADAMTS5 prevents cartilage degradation in a murine model of osteoarthritis

Sonya S. Glasson, Roger Askew, Barbara Sheppard, Brenda Carito, Tracey Blanchet, Hak-Ling Ma, Carl R. Flannery, Diane Peluso, Kim Kanki, Zhiyong Yang, Manas K. Majumdar & Elisabeth A. Morris

Nature 434, 644–648 (2005)

It has been drawn to *Nature*'s attention that E.A.M. and S.S.G. filed a patent application relevant to this work (patent number WO 2006/0004066) under the Patent Cooperation Treaty in 2004, which should therefore have been declared as a competing financial interest.

ERRATUM

doi:10.1038/nature05581

Chronic polyarthritis caused by mammalian DNA that escapes from degradation in macrophages

Kohki Kawane, Mayumi Ohtani, Keiko Miwa, Takuji Kizawa, Yoshiyuki Kanbara, Yoshichika Yoshioka, Hideki Yoshikawa & Shigekazu Nagata

Nature 443, 998–1002 (2006)

In this Letter, the units of the *y*-axis for Fig. 3d should be picograms, not micrograms. The label should read 'TNF-α in serum (pg ml^{-1})'.

CORRIGENDUM

doi:10.1038/nature05639

A Mesozoic gliding mammal from northeastern China

Jin Meng, Yaoming Hu, Yuanqing Wang, Xiaolin Wang & Chuankui Li

Nature 444, 889–893 (2006)

Of the binomen *Volaticotherium antiquus* established for a Mesozoic gliding mammal discovered from northeastern China, the termination of the trivial name should have a neuter suffix to agree with the gender of the generic name[1]. The species name is therefore corrected to *Volaticotherium antiquum*.

1. International Commission on Zoological Nomenclature. *International Code of Zoological Nomenclature* 4th edn (International Trust for Zoological Nomenclature, The Natural History Museum, London, 1999).

ERRATUM

doi:10.1038/nature05646

Fish can infer social rank by observation alone

Logan Grosenick, Tricia S. Clement & Russell D. Fernald

Nature 445, 429–432 (2007)

In the print version of this Letter, the significance level for hypothesis testing was incorrect for the 'first-choice' results (line 19 on page 430) and should be 0.05 instead of 0.01. The online PDF and HTML versions are correct.

ADDENDUM

doi:10.1038/nature05607

Artificial 'spin ice' in a geometrically frustrated lattice of nanoscale ferromagnetic islands

R. F. Wang, C. Nisoli, R. S. Freitas, J. Li, W. McConville, B. J. Cooley, M. S. Lund, N. Samarth, C. Leighton, V. H. Crespi & P. Schiffer

Nature 439, 303–306 (2006)

During the field treatment of the samples prior to measurement, the magnetic field was switched in polarity with each step down in magnitude while the sample was being rotated within the magnetic field. A more detailed description of the field treatment can be found in ref. 1.

1. Wang, R. F, *et al.* Demagnetization protocols for frustrated interacting nanomagnet arrays. *J. Appl. Phys.* (in the press); preprint at ⟨http://arxiv.org/abs/cond-mat/0702084⟩ (2007).

LETTERS

Feedback inhibition of calcineurin and Ras by a dual inhibitory protein Carabin

Fan Pan[1], Luo Sun[1]†, David B. Kardian[4], Katharine A. Whartenby[4], Drew M. Pardoll[4] & Jun O. Liu[1,2,3]

Feedback regulation of adaptive immunity is a fundamental mechanism for controlling the overall output of different signal transduction pathways, including that mediated by the T-cell antigen receptor (TCR)[1]. Calcineurin[2–4] and Ras[5–7] are known to have essential functions during T-cell activation. However, how the calcineurin signalling pathway is terminated in the process is still largely unknown. Although several endogenous inhibitors of calcineurin have been reported[8–13], none fulfils the criteria of a feedback inhibitor, as their expression is not responsive to TCR signalling. Here we identify an endogenous inhibitor of calcineurin, named Carabin, which also inhibits the Ras signalling pathway through its intrinsic Ras GTPase-activating protein (GAP) activity. Expression of Carabin is upregulated on TCR signalling in a manner that is sensitive to inhibitors of calcineurin, indicating that Carabin constitutes part of a negative regulatory loop for the intracellular TCR signalling pathway. Knockdown of Carabin by short interfering RNA led to a significant enhancement of interleukin-2 production by antigen-specific T cells *in vitro* and *in vivo*. Thus, Carabin is a negative feedback inhibitor of the calcineurin signalling pathway that also mediates crosstalk between calcineurin and Ras.

To identify new calcineurin-binding proteins we performed a yeast two-hybrid screen of a human T-cell complementary DNA library, using a truncated and catalytically inactive calcineurin mutant as bait[8]. One interacting protein identified was Cabin1, which has been extensively characterized[1,14–16]. The other calcineurin-binding protein was named Carabin on the basis of its ability to interact with both calcineurin and Ras. Carabin is unrelated to Cabin1. Human Carabin (GenBank accession number NP_940919) consists of 446 amino acid residues and is 88% identical and 91% similar in sequence to the mouse orthologue (GenBank accession number BAC30605; Fig. 1a). Carabin contains a putative Ras/Rab GAP domain at its amino terminus (residues 89–294) and a carboxy-terminal domain (residues 406–446) that mediates its interaction with calcineurin. Northern blot analysis revealed that Carabin is most abundant in spleen and peripheral blood leukocytes (Supplementary Fig. 1).

The interaction between Carabin and calcineurin was confirmed both *in vitro* and *in vivo*. A mammalian two-hybrid assay was employed to verify the interaction between Carabin and calcineurin in Jurkat T cells. Thus, the catalytic subunit of calcineurin β2 was fused to the Gal4 DNA-binding domain (pM-CNA) and the full-length Carabin was fused with the VP-16 transactivation domain (pVP-Carabin)[17]. These constructs were transfected into Jurkat T cells along with a Gal4–luciferase (Luc) reporter plasmid. As shown in Fig. 1b, calcineurin and Carabin interacted, as judged by the activation of the Gal4–Luc reporter gene. The interaction between calcineurin and Carabin was sensitive to inhibition by FK506, cyclosporin A (CsA) and the calmodulin antagonist W7 (Supplementary Fig. 2). In addition, endogenous calcineurin was also found to coimmunoprecipitate with endogenous Carabin in Jurkat T-cell lysates (Fig. 1c). We then mapped the minimal domain in Carabin that is sufficient to mediate its interaction with calcineurin by using the mammalian two-hybrid assay. It was found that the C-terminal 40-residue fragment, Carabin(406–446), is necessary and sufficient to bind to calcineurin (Fig. 1d). This was further confirmed in glutathione *S*-transferase (GST)–Carabin pull-down experiments with GST–Carabin(406–446) (Supplementary Fig. 3).

The effect of Carabin on the phosphatase activity of calcineurin was determined in an enzymatic assay with a phosphopeptide (RII) derived from the regulatory subunit of PKA (cAMP-dependent protein kinase) as a substrate[18], with recombinant Carabin produced in baculovirus-driven insect cells. Carabin inhibited the phosphatase activity of calcineurin in a dose-dependent manner with an apparent 50% inhibitory concentration (IC_{50}) of 151 nM (Fig. 2a), which is comparable to that of calcipressin[13]. The ability of Carabin to inhibit calcineurin *in vivo* was assessed by determining the effects of overexpression of Carabin on the calcium-dependent dephosphorylation of nuclear factor of activated T cells (NFAT) and its subsequent nuclear translocation. As shown previously[19], stimulation of Jurkat T cells by ionomycin led to dephosphorylation of NFAT as judged by a shift in the gel mobility of NFAT (Fig. 2b; compare lane 2 with lane 1). Ectopic expression of Carabin inhibited the dephosphorylation of NFATc2 (also known as NFAT1 or NFATp; Fig. 2b, lane 4), similar to the treatment with the calcineurin inhibitor CsA (Fig. 2b, lane 3). Consistent with its inhibitory effect on NFAT dephosphorylation was the observation that Carabin also blocked the nuclear translocation of green fluorescent protein (GFP)–NFATc4 in response to stimulation with ionomycin[20] (Fig. 2c; compare the bottom panels with the top right panel).

Calcineurin activity has been shown to be required for interleukin (IL)-2 transcription. Transient transfection with pSG-Carabin (where pSG is the parent vector) led to the inhibition of the IL-2 luciferase reporter gene activation in a dose-dependent manner (Fig. 2d). Several transcription factors have been shown to participate in the activation of the IL-2 promoter in response to TCR signalling, including NFAT, activator protein 1 (AP-1) and nuclear factor (NF)-κB[3,4]. We therefore examined the effects of Carabin on the activation of luciferase reporter genes for each transcription factor. As shown in Fig. 2e, the luciferase reporter genes for AP-1, NFAT and NF-κB, but not the constitutively active thymidine kinase promoter, were all inhibited on overexpression of Carabin. It was apparent that Carabin had more pronounced inhibitory effects on both AP-1 and NF-κB luciferase reporters than did CsA, which only partly inhibited NF-κB luciferase and had no effect on AP-1 luciferase (Supplementary Fig. 4), indicating that inhibition of calcineurin alone cannot account for the inhibitory effect of Carabin on the transcription of IL-2.

[1]Department of Pharmacology and Molecular Sciences, [2]Solomon H. Snyder Department of Neuroscience, [3]Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. [4]Immunology and Hematopoiesis Division, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, Maryland 21231, USA. †Present address: New England Biolabs, Inc., Ipswich, Massachusetts 01938, USA.

The presence of the N-terminal Ras/Rab GAP domain in Carabin raised the possibility that this domain might also be involved in the regulation of TCR signalling by inhibiting the Ras signalling pathway. To assess this possibility we determined the effects of Carabin and its N-terminal Ras/Rab GAP domain on the activation of the IL-2 luciferase reporter in the presence of a constitutively activating Ras mutant (RasV12). Unlike full-length Carabin, overexpression of its N-terminal Ras GAP domain caused a significant but partial inhibition of the IL-2 luciferase reporter activity (Fig. 3a). Coexpression of RasV12 reversed the inhibition by the Ras/Rab GAP domain of Carabin, indicating that inhibition of the IL-2 reporter activity by the Ras/Rab GAP domain of Carabin can be attributed nearly entirely to the inhibition of Ras. Indeed, recombinant Carabin was found to possess Ras GAP activity *in vitro*, as judged by the increased hydrolysis of radiolabelled GTP by Ras in the presence of Carabin in a dose-dependent manner (Fig. 3b, lanes 3–5)[21]. On immunoprecipitation from Jurkat T-cell lysates, ectopically expressed Carabin was found to be active as a Ras GAP enzyme (Fig. 3c). The N-terminal GAP domain, but not the C-terminal domain, exhibited similar Ras GAP activity when immunoprecipitated from cell lysates (Supplementary Fig. 5).

To determine further the effect of Carabin on Ras *in vivo*, we established three stable Jurkat T-cell lines carrying an empty vector (as control), wild-type Carabin and a Carabin(R141A) mutant, which was inactive in the Ras GAP assay *in vitro* (Fig. 3b), as expected[22]. When the levels of Ras-GTP were determined with a GST–Ras-binding domain pull-down assay[23], stable expression of wild-type Carabin, but not the Carabin(R141A) mutant, abolished the appearance of Ras-GTP in comparison with control cells (Fig. 3d). Similar results were obtained when the amounts of ^{32}P-labelled GTP

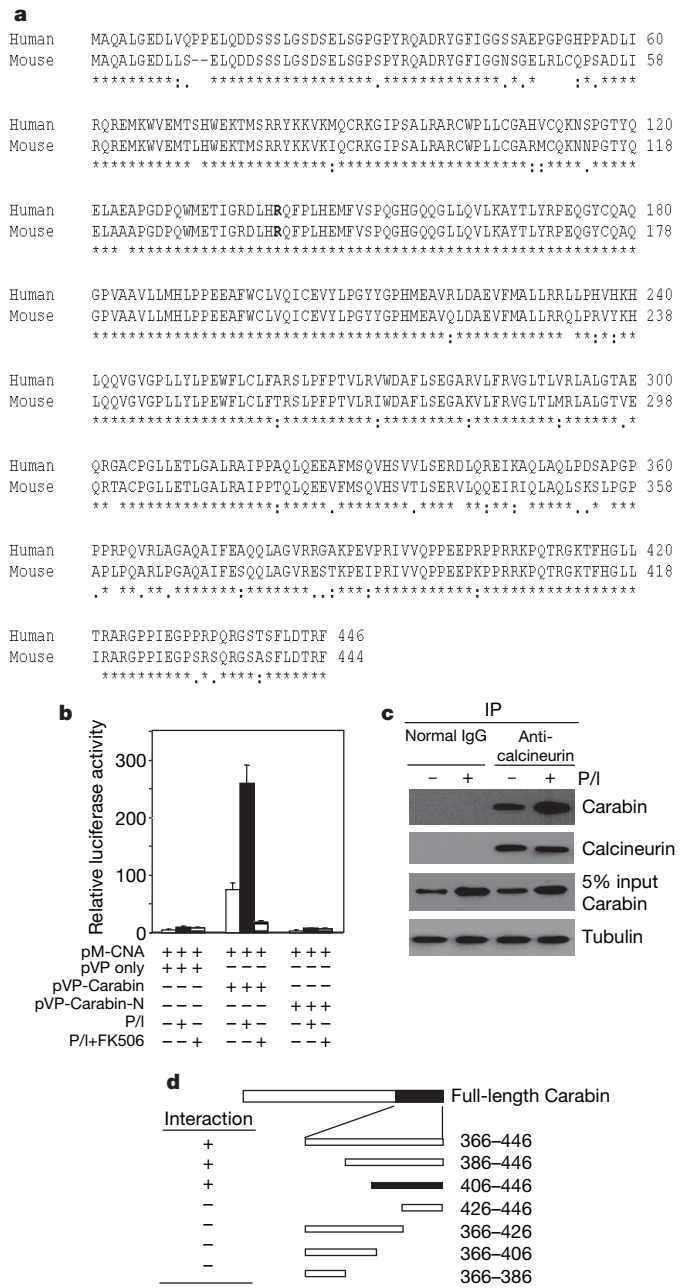


Figure 1 | Carabin interacts with calcineurin *in vitro* and *in vivo*. **a**, Sequence comparison of human and murine Carabin. Arg 141, which is critical for the Ras GAP activity, is shown in bold. **b**, Verification of the interaction between Carabin and calcineurin in a mammalian two-hybrid assay. Error bars show s.d. ($n = 3$). pVP, pVP16 vector from Clontech; P/I, PMA and ionomycin. **c**, Coimmunoprecipitation of endogenous calcineurin and Carabin. Calcineurin was immunoprecipitated followed by western blotting with anti-Carabin and anti-calcineurin antibodies. **d**, Mapping of the minimal calcineurin-binding domain of Carabin with a mammalian two-hybrid system.

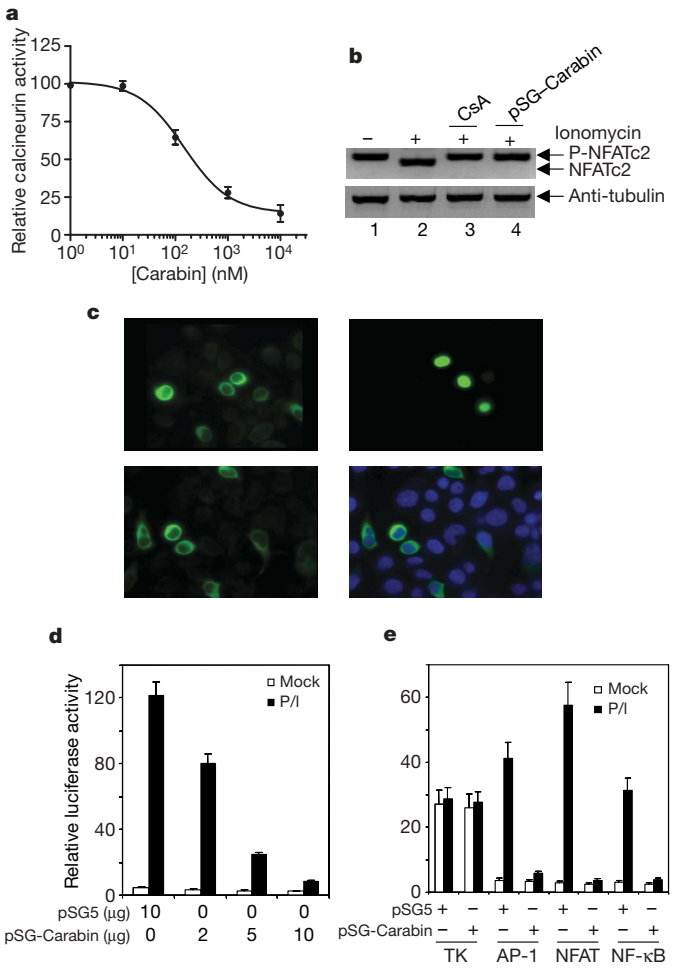


Figure 2 | Carabin inhibits calcineurin activity *in vitro* and *in vivo*. **a**, Inhibition of the phosphatase activity of calcineurin (9 nM) by recombinant Carabin *in vitro* with the use of phospho-RII (150 nM) peptide as a substrate. From the curve, $IC_{50} = 151 \pm 1.4$ nM. **b**, Carabin inhibits the dephosphorylation of NFAT on stimulation. **c**, Carabin blocks the nuclear translocation of GFP–NFATc4(1–460). Top left, unstimulated; top right, stimulated with ionomycin; bottom left, co-expression with Carabin; bottom right, as bottom left with 4′,6-diamidino-2-phenylindole staining. **d**, Overexpression of Carabin inhibits IL-2 luciferase reporter activity. **e**, Carabin inhibits the activation of AP-1–Luc, NFAT–Luc and NF-κB–Luc reporter genes. TK, thymidine kinase. Error bars show s.d. ($n = 3$).

bound to Ras were assayed on overexpression of either wild-type Carabin or the Carabin(R141A) mutant (Supplementary Fig. 6). Taken together, these results indicate that the Carabin protein possesses Ras GAP activity *in vivo*.

It has been shown that Ras activation on TCR signalling leads to the phosphorylation of both extracellular signal-regulated kinase (ERK)1 and ERK2 (ref. 24). We therefore examined the kinetics of phosphorylation of ERKs in response to stimulation with 12-*O*-tetradecanoylphorbol-13-acetate and ionomycin in a stable Jurkat T-cell line expressing the N-terminal Ras GAP domain (Carabin-N). On activation, ERK proteins underwent rapid phosphorylation, peaking at about 10 min, followed by gradual dephosphorylation as a result of the induction of feedback inhibitory mitogen-activated protein (MAP) kinase phosphatases (Fig. 3e)[1]. In comparison with control cells (−), there was a significant delay in the onset of phosphorylation of ERK1/2 (5 min versus 1 min), and an accelerated dephosphorylation when Carabin-N was overexpressed (+), with ERK1 being completely dephosphorylated by 30 min (as opposed to more than 60 min). The activation of two other related members of the MAP kinase superfamily, p38 and c-Jun N-terminal kinase, were unaffected by Carabin-N (Supplementary Fig. 7), underscoring the high specificity of Carabin for the Ras–MAP-kinase signalling pathway.

The Carabin gene is located on chromosomes 17 and 11 in the mouse and human genomes, respectively. An examination of the proximal promoter region revealed the presence of multiple consensus NFAT-binding sites (data not shown), which prompted us to speculate that the expression of Carabin might be regulated by the calcineurin signalling pathway. Indeed, the messenger RNA of Carabin is induced in a time-dependent manner on activation of primary T cells by anti-CD3/anti-CD28 (Fig. 4a). The induction of Carabin is inhibited by CsA, at both the mRNA (Fig. 4a) and protein (Supplementary Fig. 8) levels, indicating that calcineurin is required for the activation of transcription of Carabin. The dependence of Carabin expression on the activation of the calcineurin signalling pathway indicates that Carabin might serve as a feedback inhibitor of calcineurin.

To assess the role of Carabin as a negative feedback inhibitor during TCR signalling in primary T cells, we employed lentivirus-mediated short interfering RNA (siRNA) to knock down Carabin in primary human T cells[16] and determined the effect on IL-2 production. Transduction of primary human T cells with lentivirus harbouring shRNA targeting Carabin (pFUP2-siCarabin) led to a significant decrease in the amount of Carabin protein (Supplementary Fig. 9). Downregulation of Carabin expression caused a more than fivefold increase in the amount of IL-2 produced and secreted into the culture medium in comparison with control T cells (Fig. 4b), supporting the notion that Carabin has a negative regulatory function during TCR signalling.

Last, we examined the role of Carabin in antigen-specific T-cell activation *in vivo* using haemagglutinin (HA)-specific $CD4^+$

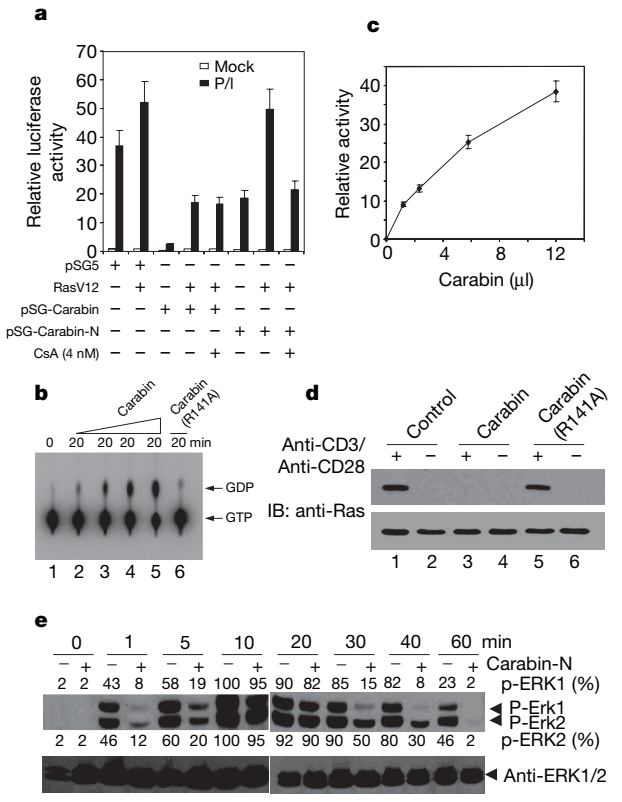


Figure 3 | Carabin possesses Ras GAP activity and decreases ERK1/2 phosphorylation on T-cell activation. **a**, Constitutively active RasV12 reverses the inhibition of IL-2 Luc reporter by full-length Carabin or its N-terminal Ras GAP domain. Relative luciferase activity is normalized to β-galactosidase activity. **b**, Carabin stimulates GTP hydrolysis by Ras as determined by thin-layer chromatography. The amounts of recombinant Carabin added were 0, 10, 20 and 50 nM in lanes 2–5, and 50 nM (lane 6). **c**, Immunoprecipitated Carabin stimulates GTP hydrolysis of Ras. The basal GTPase activity of Ras was arbitrarily set as 1. **d**, Carabin inhibits the accumulation of Ras-GTP in Jurkat T cells on stimulation with anti-CD3 and CD28. Results for bound Ras (Ras-GTP) are shown in the upper panel and those for total Ras in the bottom panel. IB, immunoblot. **e**, Effects of Carabin on phosphorylation of ERK1/2 in Jurkat T cells on stimulation. Notation: +, cells stably expressing Carabin; −, control cells. The percentages of phospho-ERK1/2 were normalized against the total ERK1/2 proteins in the corresponding lane. Error bars show s.d. ($n = 3$).

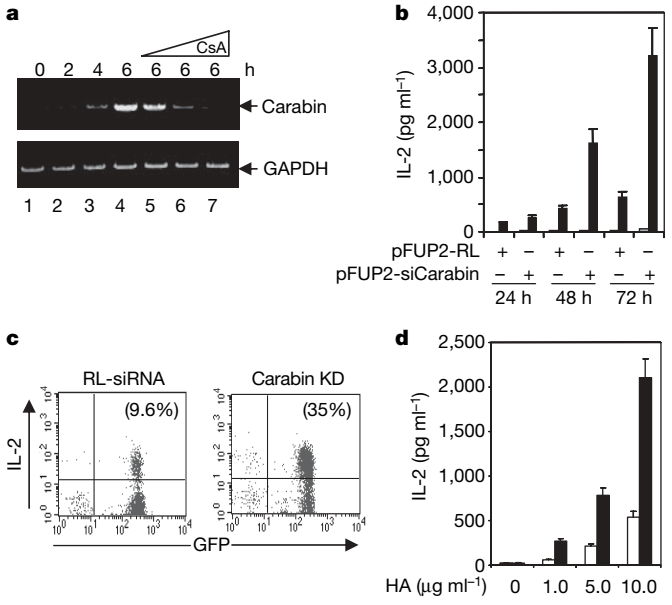


Figure 4 | Carabin expression is inducible in primary human $CD4^+$ T cells, and knockdown of Carabin causes upregulation of IL-2 production. **a**, Carabin mRNA expression was inducible in human $CD4^+$ naive T cells on stimulation with anti-CD3 and CD28, which is inhibited by CsA at 0 nM (lane 4), 10 nM (lane 5), 50 nM (lane 6) and 200 nM (lane 7). **b**, Upregulation of IL-2 production in primary human T cells on Carabin knockdown. Open columns, mock; filled columns, anti-CD3/CD28. RL, *Renilla* luciferase. **c**, Naive GFP^+CD4^+ T cells were stimulated with antigen-presenting cells/HA (10 μg/ml) for 24 h, followed by FACS analysis of intracellular IL-2. KD, knockdown. **d**, GFP^+CD4^+ T cells were stimulated with the indicated amount of HA peptide, together with APC for 24 h, and the IL-2 in the culture medium was determined by enzyme-linked immunosorbent assay. Open columns, RL-siRNA; filled columns, Carabin knockdown. Error bars show s.d. ($n = 3$).

TCR-transgenic mice (clone 6.5)[25]. In response to stimulation with HA peptide presented by irradiated syngeneic spleen cells, Carabin mRNA underwent an increase and then a decrease during a 48-h stimulation period, peaking at about 12 h (Supplementary Fig. 10). When murine Carabin was knocked down with lentivirus-mediated siRNA, there was a significant enhancement of IL-2 mRNA expression on stimulation with HA peptide (Supplementary Fig. 11). An increase in Ras-GTP was observed in murine T cells on knocking down Carabin (Supplementary Fig. 12). Moreover, a lower concentration of ionomycin was required to cause complete dephosphorylation of NFATc2 in Carabin knockdown T cells than in control cells (Supplementary Fig. 13). We then used the same siRNA lentivirus to transduce bone-marrow-derived haematopoietic stem-progenitor cells (HSC) from clone 6.5 TCR-transgenic mice. After transplantation into lethally irradiated recipients, $CD4^+$ T cells differentiating from transduced HSC will be HA-specific and have knocked down Carabin. Progeny of transduced HSC are marked by GFP. T cells were isolated from the thymus, spleen and lymph nodes of the reconstituted recipients after 10 weeks. There were no gross defects in the cellularity of TCR^+, $CD4^+$ or $CD8^+$ T cells on Carabin knockdown (Supplementary Fig. 14), indicating that Carabin is not involved in T-cell development. This is somewhat surprising given that both calcineurin and Ras have been shown to be important in T-cell development[26,27]. A nearly complete knockdown of Carabin was confirmed in GFP^+CD4^+ T cells (Supplementary Fig. 15). Upon stimulation with HA peptide, an increase in proliferation was observed in Carabin knockdown T cells (Supplementary Fig. 16). Similarly to human T cells, knockdown of Carabin also led to a significant increase in the production of intracellular (Fig. 4c) and secreted IL-2 (Fig. 4d).

Calcineurin and Ras have been shown to act synergistically to activate the transcription of IL-2 and other cytokine genes through the interaction of their respective downstream transcription factor targets, NFAT and AP-1 (ref. 6). Carabin is a negative feedback inhibitor for the calcineurin–NFAT signalling pathway that not only inhibits calcineurin but also blocks the Ras pathway, connecting calcium signalling with the downregulation of Ras. The concurrent inhibition of both calcineurin and Ras pathways by Carabin might be important in the attenuation of TCR signalling and T-cell activation. Given the prevalence of both the calcineurin and Ras pathways in other biological systems, from cell proliferation to differentiation, it is likely that Carabin is also important in other cellular processes.

METHODS

Detailed methods are given in Supplementary Information.

Coimmunoprecipitation. Jurkat T cells (2×10^7) were stimulated with 12-*O*-tetradecanoylphorbol-13-acetate (40 nM) and ionomycin (1 μM) or left untreated for 6 h before the cells were harvested and lysed in 1 ml of lysis buffer (20 mM Tris-HCl pH 7.4, 150 mM NaCl, 0.5% Nonidet P40, 0.5 mM EDTA, 10 μg ml^{-1} aprotinin, 10 μg ml^{-1} leupeptin, 1 mM phenylmethylsulphonyl fluoride and 0.5 × protease inhibitor cocktail (Sigma)). Cells were disrupted by repeated pipetting and left on ice for 10 min, followed by centrifugation at 13,000*g* for 10 min at 4 °C. For immunoprecipitation, 10 μg of antibodies or control IgG was added to the clarified supernatants and incubated with end-over-end mixing for 2 h at 4 °C. Protein A/G–Sepharose (Santa Cruz) was then added and mixing was continued for a further 1 h. The Sepharose beads were washed three times, each with 1 ml of lysis buffer, before the bound proteins were subjected to SDS–polyacrylamide-gel electrophoresis and western blot analyses.

Transduction of primary human and murine $CD4^+$ T cells with lentivirus. Cells were mixed with virus supernatants (multiplicity of infection = 5) in the presence of Polybrene (8 μg ml^{-1}) in a 5-ml tube, followed by the addition of 10 mM HEPES and centrifugation at 2,000*g* for 3 h at 30 °C. After incubation with viral supernatant for 10 h, cells were washed and incubated for 12–16 h in fresh medium containing IL-7, followed by another cycle of transduction. Cells were washed and plated at a density of 10^6 cells ml^{-1} in the presence of IL-7 as a T-cell survival factor. A fraction of cells (10^5) from each group was analysed by fluorescence-activated cell sorting to determine the efficiency of transduction (about 90%) by monitoring GFP expression 60 h after transduction. The remaining cells were stimulated as indicated.

Received 9 October; accepted 24 November 2006.
Published online 17 January 2007.

1. Liu, J. O. The yins of T cell activation. *Sci. STKE* **2005,** re1 (2005).
2. Liu, J. *et al.* Calcineurin is a common target of cyclophilin–cyclosporin A and FKBP–FK506 complexes. *Cell* **66,** 807–815 (1991).
3. Crabtree, G. R. & Clipstone, N. A. Signal transmission between the plasma membrane and nucleus of T lymphocytes. *Annu. Rev. Biochem.* **63,** 1045–1083 (1994).
4. Rao, A., Luo, C. & Hogan, P. G. Transcription factors of the NFAT family: regulation and function. *Annu. Rev. Immunol.* **15,** 707–747 (1997).
5. Downward, J., Graves, J. D., Warne, P. H., Rayter, S. & Cantrell, D. A. Stimulation of p21ras upon T-cell activation. *Nature* **346,** 719–723 (1990).
6. Woodrow, M., Clipstone, N. A. & Cantrell, D. p21ras and calcineurin synergize to regulate the nuclear factor of activated T cells. *J. Exp. Med.* **178,** 1517–1522 (1993).
7. Cantrell, D. A. GTPases and T cell activation. *Immunol. Rev.* **192,** 122–130 (2003).
8. Sun, L. *et al.* Cabin 1, a negative regulator for calcineurin signaling in T lymphocytes. *Immunity* **8,** 703–711 (1998).
9. Lai, M. M., Burnett, P. E., Wolosker, H., Blackshaw, S. & Snyder, S. H. Cain, a novel physiologic protein inhibitor of calcineurin. *J. Biol. Chem.* **273,** 18325–18331 (1998).
10. Rothermel, B. *et al.* A protein encoded within the Down syndrome critical region is enriched in striated muscles and inhibits calcineurin signaling. *J. Biol. Chem.* **275,** 8719–8725 (2000).
11. Kingsbury, T. J. & Cunningham, K. W. A conserved family of calcineurin regulators. *Genes Dev.* **14,** 1595–1604 (2000).
12. Gorlach, J. *et al.* Identification and characterization of a highly conserved calcineurin binding protein, CBP1/calcipressin, in *Cryptococcus neoformans*. *EMBO J.* **19,** 3618–3629 (2000).
13. Ryeom, S., Greenwald, R. J., Sharpe, A. H. & McKeon, F. The threshold pattern of calcineurin-dependent gene expression is altered by loss of the endogenous inhibitor calcipressin. *Nature Immunol.* **4,** 874–881 (2003).
14. Youn, H.-D., Sun, L., Prywes, R. & Liu, J. O. Apoptosis of T lymphocytes mediated by calcium-induced release of the transcription factor MEF2. *Science* **286,** 790–793 (1999).
15. Han, A. *et al.* Sequence-specific recruitment of transcriptional co-repressor Cabin1 by myocyte enhancer factor-2. *Nature* **422,** 730–734 (2003).
16. Pan, F., Means, A. R. & Liu, J. O. Calmodulin-dependent protein kinase IV regulates nuclear export of Cabin1 during T-cell activation. *EMBO J.* **24,** 2104–2113 (2005).
17. Fearon, E. R. *et al.* Karyoplasmic interaction selection strategy: a general strategy to detect protein–protein interactions in mammalian cells. *Proc. Natl Acad. Sci. USA* **89,** 7958–7962 (1992).
18. Manalan, A. S. & Klee, C. B. Activation of calcineurin by limited proteolysis. *Proc. Natl Acad. Sci. USA* **80,** 4291–4295 (1983).
19. Shaw, K. T. *et al.* Immunosuppressive drugs prevent a rapid dephosphorylation of transcription factor NFAT1 in stimulated immune cells. *Proc. Natl Acad. Sci. USA* **92,** 11205–11209 (1995).
20. Shibasaki, F., Price, E. R., Milan, D. & McKeon, F. Role of kinases and the phosphatase calcineurin in the nuclear shuttling of transcription factor NF-AT4. *Nature* **382,** 370–373 (1996).
21. Anderson, D. H. & Chamberlain, M. D. Assay and stimulation of the Rab5 GTPase by the p85 α subunit of phosphatidylinositol 3-kinase. *Methods Enzymol.* **403,** 552–561 (2005).
22. Albert, S., Will, E. & Gallwitz, D. Identification of the catalytic domains and their functionally critical arginine residues of two yeast GTPase-activating proteins specific for Ypt/Rab transport GTPases. *EMBO J.* **18,** 5216–5225 (1999).
23. Taylor, S. J. & Shalloway, D. Cell cycle-dependent activation of Ras. *Curr. Biol.* **6,** 1621–1627 (1996).
24. Leevers, S. J. & Marshall, C. J. Activation of extracellular signal-regulated kinase, ERK2, by p21ras oncoprotein. *EMBO J.* **11,** 569–574 (1992).
25. Cui, Y. *et al.* Immunotherapy of established tumors using bone marrow transplantation with antigen gene–modified hematopoietic stem cells. *Nature Med.* **9,** 952–958 (2003).
26. Neilson, J. R., Winslow, M. M., Hur, E. M. & Crabtree, G. R. Calcineurin B1 is essential for positive but not negative selection during thymocyte development. *Immunity* **20,** 255–266 (2004).
27. Fischer, A. M., Katayama, C. D., Pages, G., Pouyssegur, J. & Hedrick, S. M. The role of Erk1 and Erk2 in multiple stages of T cell development. *Immunity* **23,** 431–443 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank members of the Liu laboratory, H. Yu and G. Zhou for helpful discussions, and F. McKeon, A. Rao and D. Shalloway for plasmids. This work was supported by funds from the Department of Pharmacology, Johns Hopkins School of Medicine and the Keck Foundation.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.O.L. (joliu@jhu.edu).

LETTERS

Escape from HER-family tyrosine kinase inhibitor therapy by the kinase-inactive HER3

Natalia V. Sergina[1], Megan Rausch[1], Donghui Wang[1], Jimmy Blair[2], Byron Hann[1], Kevan M. Shokat[2] & Mark M. Moasser[1]

Oncogenic tyrosine kinases have proved to be promising targets for the development of highly effective anticancer drugs. However, tyrosine kinase inhibitors (TKIs) against the human epidermal growth factor receptor (HER) family show only limited activity against HER2-driven breast cancers, despite effective inhibition of epidermal growth factor receptor (EGFR) and HER2 *in vivo*[1–8]. The reasons for this are unclear. Signalling in *trans* is a key feature of this multimember family and the critically important phosphatidylinositol-3-OH kinase (PI(3)K)/Akt pathway is driven predominantly through transphosphorylation of the kinase-inactive HER3 (refs 9, 10). Here we show that HER3 and consequently PI(3)K/Akt signalling evade inhibition by current HER-family TKIs *in vitro* and in tumours *in vivo*. This is due to a compensatory shift in the HER3 phosphorylation–dephosphorylation equilibrium, driven by increased membrane HER3 expression driving the phosphorylation reaction and by reduced HER3 phosphatase activity impeding the dephosphorylation reaction. These compensatory changes are driven by Akt-mediated negative-feedback signalling. Although HER3 is not a direct target of TKIs, HER3 substrate resistance undermines their efficacy and has thus far gone undetected. The experimental abrogation of HER3 resistance by small interfering RNA knockdown restores potent pro-apoptotic activity to otherwise cytostatic HER TKIs, re-affirming the oncogene-addicted nature of HER2-driven tumours and the therapeutic promise of this oncoprotein target. However, because HER3 signalling is buffered against an incomplete inhibition of HER2 kinase, much more potent TKIs or combination strategies are required to silence oncogenic HER2 signalling effectively. The biologic marker with which to assess the efficacy of HER TKIs should be the transphosphorylation of HER3 rather than autophosphorylation.

Selective inhibitors of the ABL tyrosine kinase are effective in putting nearly all patients with BCR-ABL-driven leukaemia in chronic phase into complete remission[11]. This proof of concept brings new hope to the treatment of other tyrosine-kinase-driven cancers. Another important tyrosine kinase family is the HER family, consisting of EGFR, HER2, HER3 and HER4. A subset of breast cancers are driven by overactive EGFR or HER2 tyrosine kinases and an abundance of data from *in vitro* and mouse models suggests that continued activity of these tyrosine kinases drives cancer progression[12–14]. While the complexities of this tyrosine kinase family are not yet fully understood, their oncogenic signalling functions should, in theory, be amenable to silencing by TKIs. Several orally bioavailable HER-family TKIs are in preclinical and clinical development. Although in *in vitro* biochemical assays these agents differ in their relative activities against individual HER kinase family members, in cell-based assays they are effective at inhibiting both EGFR and HER2 and equally effective at suppressing the growth of EGFR- and HER2-driven tumour cells[15–19]. They are also effective at inhibiting EGFR and HER2 phosphorylation in patients' tissues and tumours[5–8]. But these agents show very limited clinical anti-tumour activity[1–5]. Their clinical development to this point has been driven largely by the detection of modest delays in tumour progression. The failure to reverse cancer progression despite an apparent inhibition of HER kinase function has created an enigma in the concept of TKI cancer therapy that we have been exploring.

It is through heterodimerization and transphosphorylation that the HER family performs its signalling functions. Importantly, downstream PI(3)K/Akt pathway signalling is predominantly mediated through the transphosphorylation of the kinase-inactive member HER3 (refs 9, 10). We have previously reported that sensitivity to HER-family TKI therapy correlates with the inhibition of PI(3)K/Akt pathway signalling[15,20]. We and others have also reported that failure to inhibit PI(3)K/Akt signalling leads to TKI resistance[20–22]. But in contrast to reports from *in vitro* models, Akt activity is not inhibited in most patients on HER TKI therapy[5,6,8]. This led us to look more closely at the inhibition of PI(3)K/Akt signalling.

To investigate this discrepancy, we studied the durability of Akt inhibition by TKI, with surprising results. Although gefitinib inhibits Akt signalling in HER2-driven cancer cells (as previously reported[15]) this inhibition is not durable. Akt signalling resumes after a transient inhibition despite continued drug therapy (Fig. 1a, b). In light of this finding, we looked at the broader HER-family signalling activities over a period of 96 h after continuous exposure of BT474 breast cancer cells to gefitinib at concentrations that non-selectively inhibit EGFR and HER2. TKI treatment effects a sustained inhibition of EGFR and HER2 phosphorylation and a durable inhibition of downstream MAPK (mitogen-activated protein kinase) and JNK (Jun N-terminal kinase) pathway signalling (Fig. 1a). However, dephosphorylation of the kinase-inactive family member HER3 is transient. HER3 signalling resumes and persists despite continued drug exposure and effective suppression of EGFR and HER2 (Fig. 1a, b). The reactivation of HER3 signalling explains the reactivation of Akt signalling, because HER3 is the principal HER-family member that binds PI(3)K and drives Akt signalling in these tumours[9,10]. TKI-refractory Akt signalling remains sensitive to PI(3)K inhibitors as expected (not shown).

These time-dependent findings are not due to drug degradation because the drug is replenished daily in these studies and HER3/Akt signalling resumes despite repeatedly refreshing drug supply up to and beyond the point of resumption of Akt signalling (not shown). There is no significant expression of HER4 before or after drug treatment in these cells (data not shown). These findings are not unique to BT474 and SkBr3 cells and have been confirmed in other HER2-overexpressing breast cancer cells, including MDA-453, AU565, MDA-361 and HCC1954 cells (Supplementary Fig. 1). These findings

[1]Department of Medicine, [2]Department of Cellular and Molecular Pharmacology, University of California, San Francisco 94143, USA.

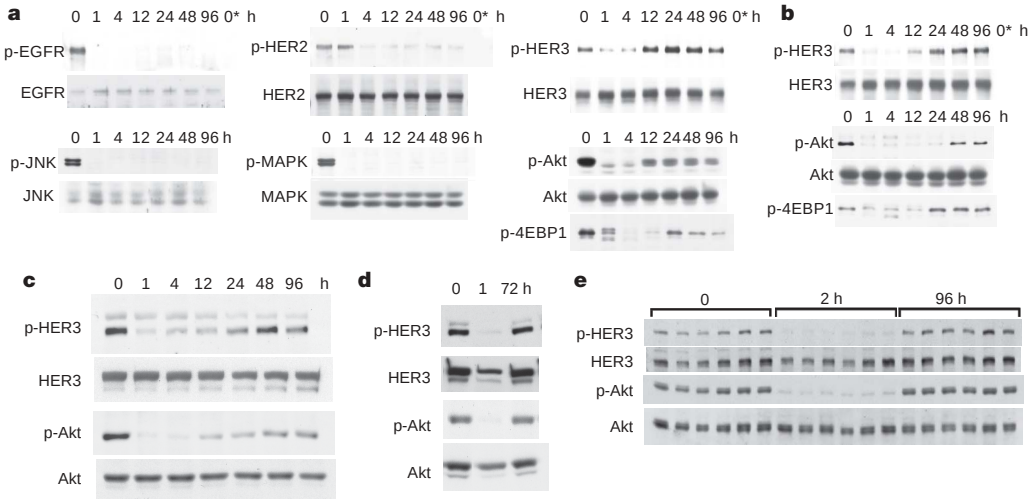


Figure 1 | HER TKIs fail to induce sustained inhibition of HER3 signalling in HER2-driven breast cancer cells. **a**, BT474 cells were treated with 5 μM gefitinib for the indicated times and assayed for expression and phosphorylation of the indicated proteins. Lane 0* is an IgG immunoprecipitation control. **b**, Data of interest shown from the identical experiment done in SkBr3 cells. **c**, SKBr3 cells were treated with 5 μM erlotinib for the indicated times and lysates immunoblotted with the indicated antibodies. **d**, SKBr3 cells were treated with 20 μM AG825 for the indicated times, with the drug being replenished 1 h before the 72 h harvest. **e**, Mice bearing HCC1569 human breast cancer xenografts were treated with gefitinib at 150 mg kg^{-1} once daily. Animals were killed 2 h after the first dose or 2 h after the fourth dose (96 h total treatment) and tumours were rapidly harvested for immunoblotting with the indicated antibodies. The six lanes corresponding to each time point represent six different animals. p, phosphorylated.

are not unique to gefitinib and are seen with other HER TKIs, including agents with *in vitro* selectivity profiles favouring EGFR or HER2, such as erlotinib or AG825 (Fig. 1c, d). Nor are these findings artefacts of the *in vitro* models. Treatment with gefitinib of mice that have various HER2-driven xenograft tumours similarly fails to durably suppress HER3 and Akt signalling, despite a transient suppression (Fig. 1e and Supplementary Fig. 2). This is not due to ineffective drug biodistribution, because in these models gefitinib was dosed three times higher than doses known to achieve sustained xenograft tumour concentrations of above 2–4 μM and averaging 6–10 μM (ref. 23). We had previously established that inactivation of PI(3)K/Akt signalling is mechanistically linked to HER TKI sensitivity in HER-family-driven cancers, so we felt that the failure of these drugs durably to inactivate PI(3)K/Akt signalling was entirely

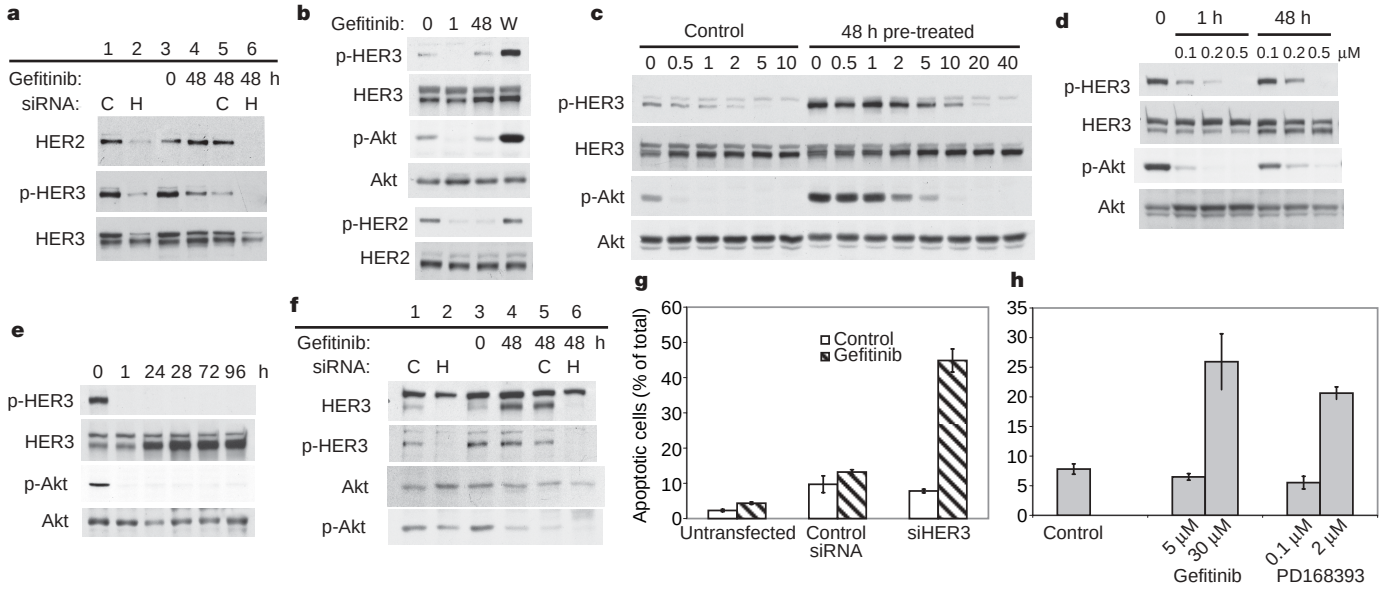


Figure 2 | Forward shift in HER3 phosphorylation–dephosphorylation equilibrium following extended HER TKI treatment. **a**, BT474 cells were transfected with anti-HER2 (H) or control (C) siRNA and harvested 64 h after transfection (lanes 1, 2). Additional arms were treated with 48 h of gefitinib untransfected (lanes 3, 4) or following siRNA transfection (lanes 5, 6). **b**, SKBr3 cells were treated with 5 μM gefitinib for 0, 1 or 48 h. Arm W was treated for 48 h, washed, and incubated in drug-free media for one more hour. **c**, SKBr3 cells were treated with the indicated concentrations of gefitinib for one hour (left side). Additional arms were treated with 5 μM gefitinib for 48 h and subsequently treated with the indicated concentrations of gefitinib for one additional hour (right side). **d**, SKBr3 cells were treated with the indicated concentrations and durations of PD168393. **e**, SKBr3 cells were treated with 2 μM PD168393 for the indicated times. **f**, SKBr3 cells were transfected with anti-HER3 (H) or control (C) siRNA and harvested four days after transfection (lanes 1, 2). Additional arms were treated with gefitinib or control in untransfected cells (lanes 3, 4) or following siRNA transfection (lanes 5, 6). **g**, In parallel with **f**, SkBr3 cells were either left untransfected, or transfected with anti-HER3 or control siRNA followed by 5 μM gefitinib or control for 48 h. Apoptotic cells were identified by their sub-G1 DNA content. **h**, SKBr3 cells were treated as indicated for 48 h. Apoptotic cells were identified by Annexin V expression. Error bars, s.d.

consistent with their limited clinical activities. Therefore we set out to study the molecular mechanism by which HER3 evades TKI therapy.

TKI-refractory HER3 phosphorylation is due to HER2, because it can be suppressed by anti-HER2 siRNA transfection (Fig. 2a). Although cross-talk between receptor families can occur, we have found no evidence for non-HER-family tyrosine kinases mediating TKI-refractory HER3 phosphorylation. The reactivation of HER3 signalling is not associated with the induction of any new tyrosine kinases and remains resistant to the broad-spectrum kinase inhibitor staurosporine (not shown). This apparent desensitization of HER3 signalling to TKIs is due to a forward shift in the equilibrium of the HER3 phosphorylation–dephosphorylation reactions, establishing a new steady-state HER3 phosphorylation level despite significant inhibition of HER2 kinase and autophosphorylation activity by TKIs. This forward shift becomes clearly evident in the form of HER3 and Akt superphosphorylation when drug inhibition is withdrawn during the new steady state (Fig. 2b). HER3 phosphorylation remains suppressible by TKI in the new steady state, but much higher concentrations are required to completely dephosphorylate HER3 because the un-inhibited HER3 phosphorylation state is significantly higher in the new steady state (Fig. 2c, compare left and right). Therefore drug-refractory HER3 phosphorylation is due to resistance at the level of the substrate HER3, and is driven by residual HER2 kinase activity. Similar characteristics apply to the more potent irreversible TKIs[24]. The irreversible TKI PD168393, which inhibits all HER-family kinases, when used at partially or near-maximal inhibitory concentrations, induces a similar desensitization of HER3 to continued drug therapy (Fig. 2d, 0.1–0.2 μM doses). However, at fully inactivating concentrations both reversible (Fig. 2c, 40 μM dose) and irreversible (Fig. 2e) TKI can durably suppress HER3 and Akt signalling.

The biological consequence of drug-refractory HER3 and Akt signalling is tumour cell survival. In fact, the anti-proliferative activity of TKIs is reversible and tumour cells resume proliferative growth after drug withdrawal. If drug-refractory HER3 signalling is averted by anti-HER3 siRNA, TKI treatment of HER2-driven cancer cells leads to apoptotic tumour cell death (Fig. 2f, g, and Supplementary Fig. 4). This is the expected outcome of effective oncoprotein inactivation, and recapitulates the apoptotic fate of oncogene withdrawal seen in reversible transgenic models of HER2 tumorigenesis[14]. Sustained inhibition of HER3 signalling using TKIs at their fully inactivating doses (from Fig. 2c–e) also leads to apoptotic tumour cell death not seen with doses that allow HER3 escape (Fig. 2h).

The TKI-induced forward shift in the HER3 phosphorylation–dephosphorylation steady state is due to increased HER3 substrate concentration driving the forward reaction, and decreased phosphatase activity impeding the reverse reaction. Increased HER3 substrate concentration occurs through a significant increase in HER3 expression at the plasma membrane where the phosphorylation reaction occurs (Fig. 3a, b). Unlike HER2, which is predominantly localized to the plasma membrane, the HER3 pool is largely within intracellular compartments, but there is also some membrane expression[25]. The TKI-induced forward shift in HER3 steady-state phosphorylation is driven by HER3 relocalization to the plasma membrane and can be suppressed by inhibitors of vesicular trafficking (Fig. 3c, d). The HER3 dephosphorylation rate is also slowed after 48 h of TKI exposure (Fig. 3e). The retarded HER3 dephosphorylation rate may be due to reduced access to cytosolic protein tyrosine phosphatases as a result of altered endocytic trafficking, or it may be due to inhibition of protein tyrosine phosphatases. In support of the latter, TKI therapy increases the concentration of cellular reactive oxygen species (Fig. 3f), which are known to inhibit protein tyrosine phosphatases and are thus emerging as an important regulator of their activity[26,27]. Consistent with this, drug-refractory HER3 signalling can be suppressed by concomitant treatment with certain anti-oxidants (Fig. 3g).

The changes in steady-state HER3 signalling that evolve with TKI treatment are driven by the loss of Akt signalling and probably involve Akt-mediated negative feedback signalling. Consistent with this, HER3 signalling does not escape TKI treatment when a constitutively active Akt is transfected (Fig. 4a). Conversely, inhibition of

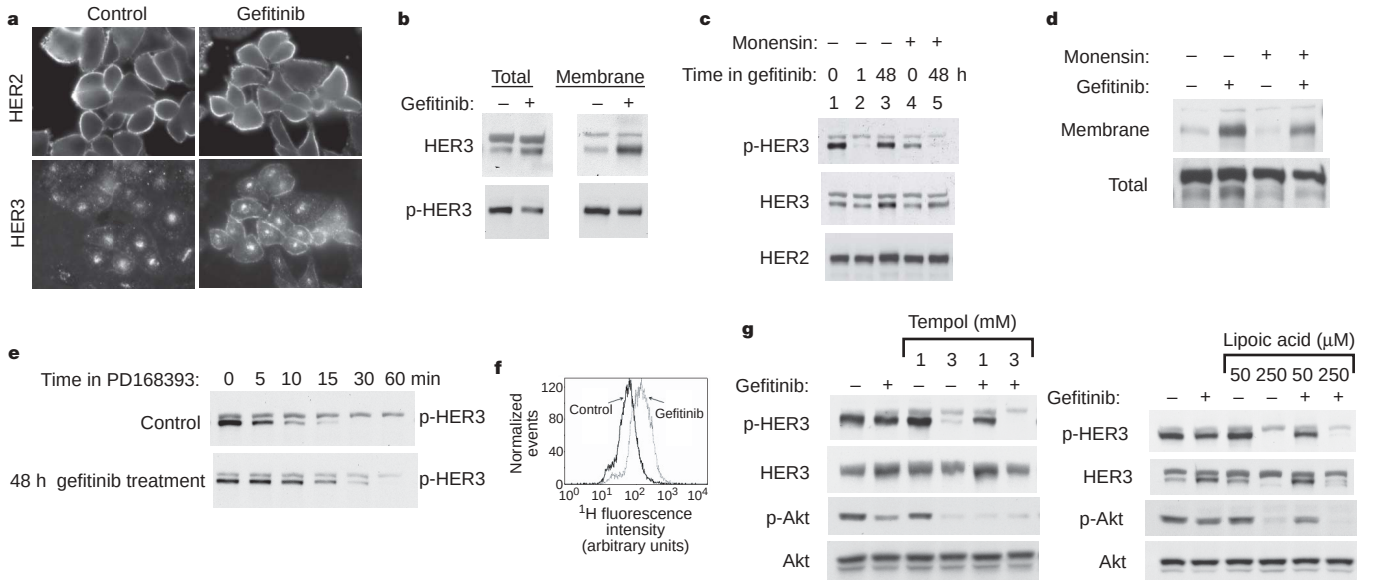


Figure 3 | Mechanism of HER3 reactivation after extended HER TKI treatment. a, SKBr3 cells were treated with 5 μM gefitinib or control for 48 h and stained with anti-HER2 or anti-HER3 antibodies for immunofluorescence microscopy. **b**, In addition, cell surface proteins in control or 48 h pre-treated cells were biotinylated, precipitated and immunoblotted as indicated. **c**, Control or gefitinib pre-treated (5 μM, 48 h) SKBr3 cells were treated with 20 μM monensin for the final 6 h and analysed by western blotting. **d**, SkBr3 cells were treated with gefitinib for 48 h and with 20 μM monensin for the final 6 h. Membrane and total HER3 was immunoblotted from cell surface proteome pulldowns (above) or total lysates (below). **e**, The dephosphorylation rate of p-HER3 following 48 h of gefitinib treatment was determined immediately after initiation of fully inactivating concentrations of the irreversible HER TKI PD168393 (2 μM). **f**, Reactive oxygen species were quantified in control or gefitinib pre-treated (5 μM, 48 h) SKBr3 cells, as described in the Methods. **g**, SKBr3 cells were treated with 5 μM gefitinib or in combination with the indicated concentrations of the anti-oxidants Tempol or α-lipoic acid for 48 h.

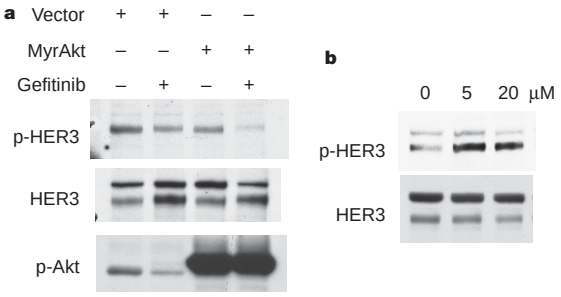


Figure 4 | Akt regulates HER3 signalling via negative feedback signalling. **a**, 2×10^6 SKBr3 cells were transiently transfected with 5 μg of pcDNA3-myristoyl-Akt plasmid or control pcDNA3 plasmid. The following day 5 μM gefitinib was added for an additional 48 h where indicated. Lysates were immunoblotted as indicated. **b**, SKBr3 cells were treated with 5 or 20 μM LY294002 for 8 h and lysates immunoblotted as indicated.

Akt signalling by a PI(3)K inhibitor leads to a compensatory increase in HER3 phosphorylation (Fig. 4b). Complete inactivation of HER2 kinase with high doses of TKIs induces the maximum feedback signalling and HER3 redistribution, but owing to the complete inactivation of HER2 kinase, HER3 signalling cannot be restored and so the feedback loop fails to rescue Akt activity.

Standard preclinical models have notoriously overestimated the clinical potential of HER TKIs, so we challenged the traditional approach to evaluating TKIs in these models. Traditionally, signalling inhibitors are thought to have a continuous suppressive effect through rapid and sustained inhibition of their direct molecular targets and downstream signalling events. This notion of drug therapy may be too simplistic. Clearly, continuous exposure to a growth factor stimulus does not produce continuous high-output downstream signalling. Rather, it leads to a sequence of signalling events, programmed by negative and positive feedback signalling, until establishment of a new steady state in the presence of continued stimulus. We find here that continuous exposure to TKIs similarly leads to a sequence of signalling events that manifest over time, until a new steady state is reached.

With this new perspective, we report that HER3 and PI(3)K/Akt signalling is not effectively inhibited by current TKIs. In particular, the allocation of kinase and signalling functions to different members within the HER family allows the signalling substrate HER3 to restore signalling activity despite significant inhibition of HER2 kinase, in effect buffering HER3-mediated PI(3)K/Akt signalling against an incomplete loss of HER2 kinase function. This inherent signal buffering capacity allows tumour cells to evade the pro-apoptotic effects of TKIs, making such TKI therapy considerably less effective. To treat HER-driven cancers much more effectively may thus require drugs with much higher potency or drugs that completely inactivate HER kinase function. Irreversible TKIs, although more potent, are subject to similar limitations. Owing to their reactive groups and reduced selectivity, many irreversible agents cannot be delivered at completely inactivating doses. Future highly selective irreversible inhibitors may turn out to be more effective. Until highly specific and fully inactivating drugs can be designed, combination treatment strategies designed to undermine the resiliency of HER-family signalling may offer the most promising approach in the near future. In addition, inhibition of autophosphorylation activity deceptively overstates the efficacy of TKIs and is a poor *in vivo* biologic marker. A better biological marker of efficacy with which to guide future therapies should be HER3 transphosphorylation.

The signal buffering capacity endowed by the separation of kinase and signalling functions to different family members in the HER kinase family attests to an evolutionary advantage conferred by the loss of catalytic activity in the HER3 protein kinase. This may be one of the reasons why approximately 10% of human kinases appear to be catalytically inactive[28].

METHODS

Cell culture and reagents. PD168393 was synthesized as previously described[29]. Commercially available gefitinib and erlotinib were purified for *in vitro* use. Reagent sources are detailed in Supplementary Information. For immunofluorescence studies, cells grown on fibronectin-coated cover slips were treated as indicated, fixed in 4% paraformaldehyde, permeabilized, and stained with the indicated primary antibodies and fluoroscein isothiocyanate (FITC)-conjugated secondary antibodies. Cells were visualized using a Zeiss Axioplan 2 fluorescence imaging microscope.

Apoptosis. Cells were seeded at 300,000–500,000 per well in 12-well or 6-well clusters. Apoptotic cells were identified and quantified by analysis of Annexin V binding using the Annexin V-FITC apoptosis detection kit (Calbiochem) according to the manufacturer's instructions, or by their sub-G1 DNA content and quantified by fluorescence-activated cell sorting (FACS) analysis as previously described[30]. All experimental arms were done in duplicate and displayed as averages with standard of deviation (s.d.) error bars.

Transfections. Cells were seeded at a density of 300,000 cells per well in 12-well plates and transfected the following day. For siRNA transfections 100–300 nmol of siRNA (Dharmacon) was premixed with Lipofectamine2000 in Opti-MEM media and then added to each well. For plasmid transfections, 2 μg of plasmid DNA was premixed with Lipofectamine2000 in Opti-MEM media and added to wells for 6 h.

Cell surface biotinylation. Cells were chilled on ice and rinsed twice with ice-cold PBS. Freshly prepared sulpho-NHS-SS-biotin (labelling reagent) was added to the final concentration of 0.5 mg ml^{-1} in PBS. After 45 min incubation at 4 °C, cells were lysed for immunoprecipitation.

Reactive oxidation species assay. Cells were rinsed twice with PBS and incubated with 10 μM of freshly prepared H2DCFDA in phenol-red-free media for 45 min at 37 °C. Cells were then trypsinized and reactive oxygen species levels were detected by flow cytometry.

Received 14 June; accepted 22 November 2006.
Published online 7 January 2007.

1. Winer, E. P. *et al.* Phase II multicenter study to evaluate the efficacy and safety of Tarceva (Erlotinib HCl, OSI-774) in women with previously treated, locally advanced or metastatic breast cancer. *Breast Cancer Res. Treatm.* **76** (suppl. 1), abstr. 445 (2002).
2. Blackwell, K. *et al.* A phase II, open-label, multicenter study of GW572016 in patients with trastuzumab-refractory metastatic breast cancer. *Proc. Am. Soc. Clin. Oncol.* **23,** abstr. 3006 (2004).
3. Dees, E. C. *et al.* Clinical summary of 67 heavily pre-treated patients with metastatic carcinomas treated with GW572016 in a phase Ib study. *Proc. Am. Soc. Clin. Oncol.* **23,** abstr. 3188 (2004).
4. Campos, S. M. *et al.* A phase 2, single agent study of CI-1033 administered at two doses in ovarian cancer patients who failed platinum therapy. *Proc. Am. Soc. Clin. Oncol.* **23,** abstr. 5054 (2004).
5. Baselga, J. *et al.* Phase II and tumor pharmacodynamic study of gefitinib in patients with advanced breast cancer. *J. Clin. Oncol.* **23,** 5323–5333 (2005).
6. Bacus, S. S., Beresford, P. J., Yarden, Y., Spector, N. & Smith, B. The use of predicting factors and surrogate markers in patients' cancer biopsies treated with targeted antibodies to erbB receptors and erbB tyrosine kinase inhibitors. *Proc. Am. Soc. Clin. Oncol.* **22,** abstr. 3408 (2003).
7. Burris, H. A. *et al.* EGF10004: a randomized, multicenter, phase Ib study of the safety, biologic activity and clinical efficacy of the dual kinase inhibitor GW572016. *Breast Cancer Res. Treatm.* **82** (suppl. 1), abstr. 39 (2003).
8. Spector, N. L. *et al.* Study of the biologic effects of lapatinib, a reversible inhibitor of ErbB1 and ErbB2 tyrosine kinases, on tumor growth and survival pathways in patients with advanced malignancies. *J. Clin. Oncol.* **23,** 2502–2512 (2005).
9. Soltoff, S. P., Carraway, K. L. III, Prigent, S. A., Gullick, W. G. & Cantley, L. C. ErbB3 is involved in activation of phosphatidylinositol 3-kinase by epidermal growth factor. *Mol. Cell. Biol.* **14,** 3550–3558 (1994).
10. Kim, H. H., Sierke, S. L. & Koland, J. G. Epidermal growth factor-dependent association of phosphatidylinositol 3-kinase with the erbB3 gene product. *J. Biol. Chem.* **269,** 24747–24755 (1994).
11. Kantarjian, H. *et al.* Hematologic and cytogenetic responses to imatinib mesylate in chronic myelogenous leukemia. *N. Engl. J. Med.* **346,** 645–652 (2002).
12. Muller, W. J., Sinn, E., Pattengale, P. K., Wallace, R. & Leder, P. Single-step induction of mammary adenocarcinoma in transgenic mice bearing the activated *c-neu* oncogene. *Cell* **54,** 105–115 (1988).
13. Slamon, D. J. *et al.* Human breast cancer: correlation of relapse and survival with amplification of the *HER-2/neu* oncogene. *Science* **235,** 177–182 (1987).
14. Moody, S. E. *et al.* Conditional activation of Neu in the mammary epithelium of transgenic mice results in reversible pulmonary metastasis. *Cancer Cell* **2,** 451–461 (2002).
15. Moasser, M. M., Basso, A., Averbuch, S. D. & Rosen, N. The tyrosine kinase inhibitor ZD1839 ("Iressa") inhibits HER2-driven signaling and suppresses the growth of HER2-overexpressing tumor cells. *Cancer Res.* **61,** 7184–7188 (2001).

16. Moulder, S. L. *et al.* Epidermal growth factor receptor (HER1) tyrosine kinase inhibitor ZD1839 (Iressa) inhibits HER2/*neu* (*erb*B2)-overexpressing breast cancer cells *in vitro* and *in vivo*. *Cancer Res.* **61,** 8887–8895 (2001).
17. Anderson, N. G., Ahmad, T., Chan, K., Dobson, R. & Bundred, N. J. ZD 1839 (Iressa), a novel epidermal growth factor receptor (EGFR) tyrosine kinase inhibitor, potently inhibits the growth of EGFR-positive cancer cell lines with or without erbB2 overexpression. *Int. J. Cancer* **94,** 774–782 (2001).
18. Campiglio, M. *et al.* Inhibition of proliferation and induction of apoptosis in breast cancer cells by the epidermal growth factor receptor (EGFR) tyrosine kinase inhibitor ZD1839 ('Iressa') is independent of EGFR expression level. *J. Cell. Physiol.* **198,** 259–268 (2004).
19. Akita, R. W. & Sliwkowski, M. X. Preclinical studies with Erlotinib (Tarceva). *Semin. Oncol.* **30** (suppl. 7), 15–24 (2003).
20. She, Q., Solit, D., Basso, A. & Moasser, M. M. Resistance to gefitinib (ZD1839, Iressa) in PTEN null HER overexpressing tumor cells can be overcome through restoration of PTEN function or pharmacologic modulation of constitutive PI3K/Akt pathway signaling. *Clin. Cancer Res.* **9,** 4340–4346 (2003).
21. Bianco, R. *et al.* Loss of PTEN/MMAC1/TEP in EGF receptor-expressing tumor cells counteracts the antitumor action of EGFR tyrosine kinase inhibitors. *Oncogene* **22,** 2812–2822 (2003).
22. Haas-Kogan, D. A. *et al.* Epidermal growth factor receptor, protein kinase B/Akt, and glioma response to erlotinib. *J. Natl Cancer Inst.* **97,** 880–887 (2005).
23. McKillop, D. *et al.* Tumor penetration of gefitinib (Iressa), an epidermal growth factor receptor tyrosine kinase inhibitor. *Mol. Cancer Ther.* **4,** 641–649 (2005).
24. Fry, D. W. *et al.* Specific, irreversible inactivation of the epidermal growth factor receptor and erbB2, by a new class of tyrosine kinase inhibitor. *Proc. Natl Acad. Sci. USA* **95,** 12022–12027 (1998).
25. Offterdinger, M., Schofer, C., Weipoltshammer, K. & Grunt, T. W. c-erbB-3: a nuclear protein in mammary epithelial cells. *J. Cell Biol.* **157,** 929–939 (2002).
26. Meng, T. C., Fukada, T. & Tonks, N. K. Reversible oxidation and inactivation of protein tyrosine phosphatases *in vivo*. *Mol. Cell* **9,** 387–399 (2002).
27. Tonks, N. K. Redox Redux: Revisiting PTPs and the control of cell signaling. *Cell* **121,** 667–670 (2005).
28. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science* **298,** 1912–1934 (2002).
29. Tsou, H. R. *et al.* 6-Substituted-4-(3-bromophenylamino)quinazolines as putative irreversible inhibitors of the epidermal growth factor receptor (EGFR) and human epidermal growth factor receptor (HER-2) tyrosine kinases with enhanced antitumor activity. *J. Med. Chem.* **44,** 2719–2734 (2001).
30. Huron, D. R. *et al.* A novel pyridopyrimidine inhibitor of abl kinase is a picomolar inhibitor of Bcr-abl-driven K562 cells and is effective against STI571-resistant Bcr-abl mutants. *Clin. Cancer Res.* **9,** 1267–1273 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by the Susan Komen Foundation (M.M.M.), the California Breast Cancer Research Program (M.M.M.), and an NIH grant (K.M.S.). We thank D. Stokoe and F. McCormick for review of the manuscript.

Author Contributions All authors contributed to the experiments in this work. The studies were conceived by M.M.M. with additional contributions from N.V.S. and K.M.S. The paper was written by N.V.S. and M.M.M.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.M.M. (mmoasser@medicine.ucsf.edu).

LETTERS

Mitotic occupancy and lineage-specific transcriptional control of *rRNA* genes by Runx2

Daniel W. Young[1]†*, Mohammad Q. Hassan[1]*, Jitesh Pratap[1], Mario Galindo[1]†, Sayyed K. Zaidi[1], Suk-hee Lee[1], Xiaoqing Yang[1], Ronglin Xie[1], Amjad Javed[1]†, Jean M. Underwood[1], Paul Furcinitti[2], Anthony N. Imbalzano[1], Sheldon Penman[3], Jeffrey A. Nickerson[1], Martin A. Montecino[4], Jane B. Lian[1], Janet L. Stein[1], Andre J. van Wijnen[1] & Gary S. Stein[1]

Regulation of ribosomal RNA genes is a fundamental process that supports the growth of cells and is tightly coupled with cell differentiation. Although rRNA transcriptional control by RNA polymerase I (Pol I) and associated factors is well studied, the lineage-specific mechanisms governing rRNA expression remain elusive[1]. Runt-related transcription factors Runx1, Runx2 and Runx3 establish and maintain cell identity[2], and convey phenotypic information through successive cell divisions for regulatory events that determine cell cycle progression or exit in progeny cells[3]. Here we establish that mammalian Runx2 not only controls lineage commitment and cell proliferation by regulating genes transcribed by RNA Pol II, but also acts as a repressor of RNA Pol I mediated rRNA synthesis. Within the condensed mitotic chromosomes we find that Runx2 is retained in large discrete foci at nucleolar organizing regions where *rRNA* genes reside. These Runx2 chromosomal foci are associated with open chromatin, co-localize with the RNA Pol I transcription factor UBF1, and undergo transition into nucleoli at sites of rRNA synthesis during interphase. Ribosomal RNA transcription and protein synthesis are enhanced by Runx2 deficiency that results from gene ablation or RNA interference, whereas induction of Runx2 specifically and directly represses rDNA promoter activity. Runx2 forms complexes containing the RNA Pol I transcription factors UBF1 and SL1, co-occupies the *rRNA* gene promoter with these factors *in vivo*, and affects local chromatin histone modifications at rDNA regulatory regions. Thus Runx2 is a critical mechanistic link between cell fate, proliferation and growth control. Our results suggest that lineage-specific control of ribosomal biogenesis may be a fundamental function of transcription factors that govern cell fate.

Runx factors are scaffolding proteins that localize to subnuclear domains and integrate cell signals through the formation of gene promoter regulatory complexes[4,5]. A dynamic intracellular reorganization of the gene regulatory machinery takes place during mitosis. In prophase, chromosomes condense, nucleoli disassemble, the nucleus reorganizes, and transcription is silenced. We have shown the disruption of Runx2 subnuclear localization during mitosis and restoration when transcription resumes in telophase[3]. Although many transcription factors are displaced from chromosomes and/or degraded during mitosis[6–9], Runx proteins remain stable and associate with mitotic chromatin[3].

Immunofluorescence microscopy of human and mouse metaphase chromosome spreads reveals that Runx2 is localized to large foci that are equivalently positioned on sister chromatids (Supplementary Fig. 1a–d). Mitotic foci are also observed for a Runx2 mutant (ΔC) with a carboxy-terminal truncation that retains the Runt-homology DNA binding domain (Supplementary Fig. 1d), but not in a cleidocranial dysplasia related DNA-binding mutant (data not shown). These results indicate that chromosomal association is independent of Runx2 C-terminal functions and involves recruitment of Runx2 to its cognate DNA motifs. Colocalization studies with antibodies against histone modifications and DNaseI hypersensitivity assays performed on mitotic chromosomes indicate that Runx2 foci reside in regions of open chromatin (Supplementary Fig. 1e–i). This unique focal organization of the lineage-specific Runx2 protein on mitotic chromosomes has not previously been documented for an RNA Pol II transcription factor, and suggests a novel regulatory function for Runx2 during mitosis.

The size and pairwise symmetric nature of the mitotic Runx2 foci and their localization with decondensed chromatin suggest that Runx2 is clustered at gene-rich chromosomal loci. From a cytogenetic perspective, Runx2 foci localize to pericentromeric regions of human and mouse chromosomes, and are positioned on human acrocentric chromosomes (Supplementary Fig. 2) that contain hundreds of tandemly arranged ribosomal genes[10]. We therefore postulated that Runx2 binds *rRNA* genes during mitosis and may control RNA Pol I transcription. This hypothesis challenges the current model that Runx proteins determine cell fate and cell cycle progression exclusively through control of RNA Pol II transcribed genes[2,11,12]. The RNA Pol I regulatory protein upstream binding factor 1 (UBF1) binds directly to *rDNA* and to mitotic nucleolar organizing regions, the precursors to interphase nucleoli[2,13–15]. On mitotic chromosomes from both human and mouse cells Runx2 foci co-localize with UBF1 at active nucleolar organizing regions that are enriched for spatially clustered *rRNA* genes (Fig. 1 and Supplementary Figs 2, 3).

When ribosomal biogenesis resumes in interphase, the *rRNA* transcriptional regulator UBF1 concentrates at nucleolar sites of rRNA synthesis[15,16]. During interphase Runx2 exhibits a punctate distribution throughout the mammalian nucleus and a subset is localized with UBF1 within nucleoli (Fig. 1b). Immunogold electron microscopy corroborates the presence of Runx2 in nucleoli (Fig. 1c and

[1]Department of Cell Biology and Cancer Center, [2]Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, Massachusetts 01655, USA. [3]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [4]Departamento de Biologia Molecular, Facultad de Ciencias Biologicas, Universidad de Concepcion, Concepcion, Chile. †Present addresses: Novartis Institutes for BioMedical Research, Cambridge, Massachusetts 02139, USA (D.W.Y.); Program of Cellular and Molecular Biology, Institute of Biomedical Sciences (I.C.B.M.), Faculty of Medicine, University of Chile, Santiago, Chile (M.G.); Institute of Oral Health Research, School of Dentistry, University of Alabama at Birmingham, Birmingham, Alabama 35294 USA (A.J.).
*These authors contributed equally to this work.

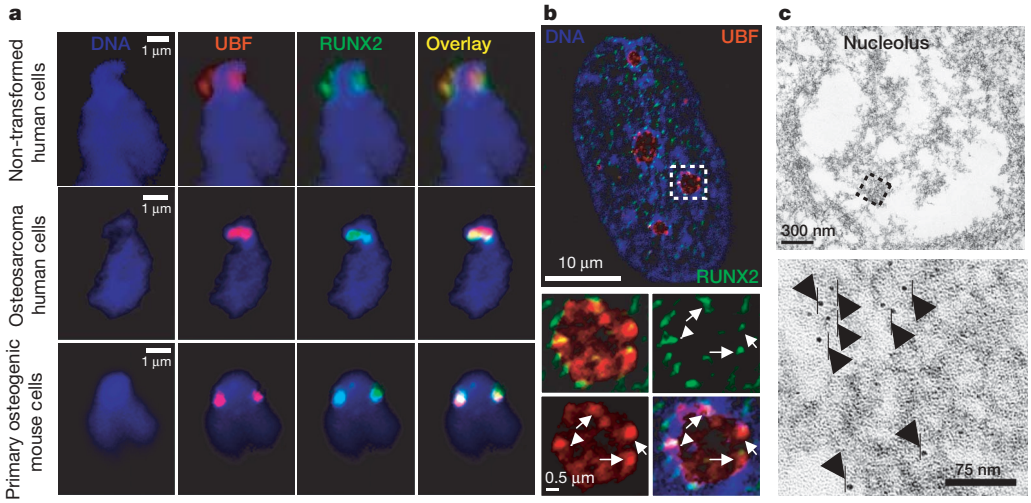


Figure 1 | Colocalization of Runx2 and UBF1 during mitosis and interphase. **a**, **b**, Immunofluorescence microscopy for Runx2 (green) and UBF1 (red) with DAPI staining (blue) and overlay for colocalization. **a**, Mitotic chromosome spreads from MCF-10A (top), Saos (middle), or primary mouse calvarial cells (bottom). **b**, Interphase Saos cell with enlarged images of the nucleolus (inset). Arrows indicate Runx2 and UBF1 and overlap. **c**, Epon section immunoelectron microscopy of Saos cells with 5 nm gold beads[29]. Top, Runx2 in the nucleolus (see Supplementary Fig. 4 for context); boxed area is shown enlarged below. Bottom, arrow heads mark gold-bead-labelled Runx2.

Supplementary Fig. 4a–e). *In situ* transcriptional run-on analysis reveals that Runx2 and UBF1 co-localize with nucleolar sites of BrUTP incorporation in human Saos cells even upon specific inhibition of RNA Pol II and III transcription with α-amanitin (Supplementary Fig. 5 and data not shown). Thus, Runx2 may have a novel and mitotically heritable role in transcriptional regulation of *rRNA* genes that is independent of its RNA Pol II related functions.

Multiple Runx binding-motifs are present throughout human, mouse and rat *rDNA* loci (Supplementary Fig. 6 and data not shown). Chromatin immunoprecipitations with interphase and mitotic cells reveal direct binding of both UBF1 and Runx2 across the *rDNA* repeat, including the transcription initiation region (Fig. 2a). Specificity is reflected by absence of Runx2 and UBF1 at the *IgH* locus and binding of Runx2, but not UBF1, to the *Runx2*

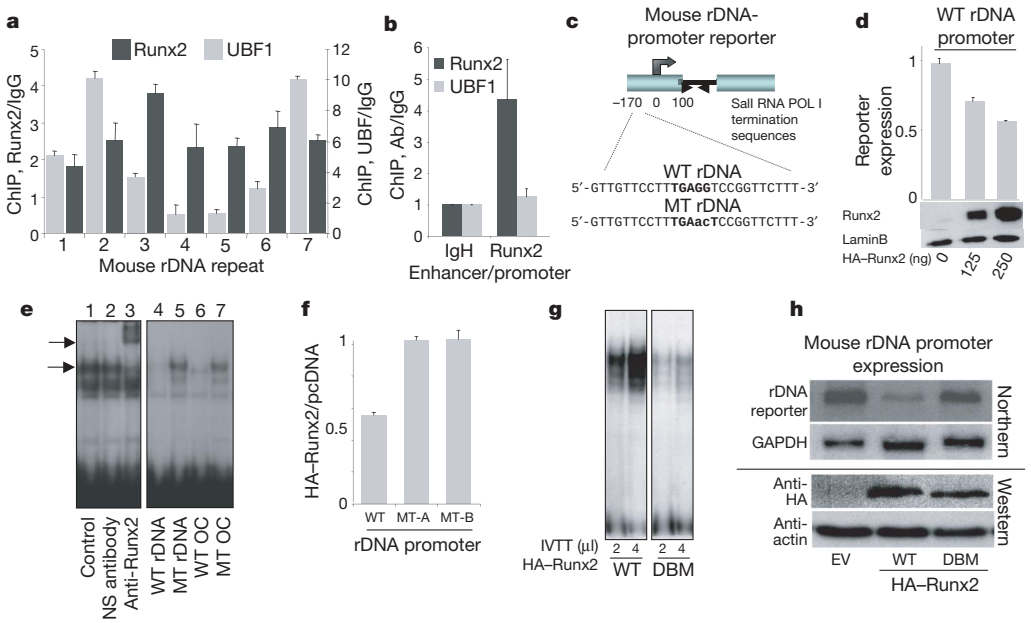


Figure 2 | Runx2 interacts with *rDNA* loci *in vivo* and directly represses rRNA transcription. **a**, **b**, ChIPs with Runx2, UBF1 and IgG antibodies show Runx2 binding to the *rDNA* repeat, the *Runx2* gene, and not the *IgH* gene in MC3T3 cells; qPCR data are normalized to non-specific genomic DNA and relative to IgG (see Supplementary Table 1 and Supplementary Figs 6 and 7). **c**, Schematic of the rDNA-promoter reporter used in **d** and **f**. Sequences of wild-type (WT) and mutant (MT) oligos used in **e** are shown with Runx binding site in bold. These sequences also correspond with WT and MT (independent clones A and B) reporters used in **f**. **d**, RT–qPCR based reporter assay showing Runx2 inhibition of rDNA transcription (see Supplementary Table 1). MC3T3 cells were transfected with Haemagglutinin (HA) tagged *Runx2* at the indicated nanogram (ng) quantities and monitored by western blot against Runx2, as a transfection control, and Lamin B, as a loading control. **e**, Electrophoretic mobility shift assays with ROS cell nuclear protein and ^{32}P-labelled rDNA oligo (see **c**). Binding of Runx2 (lower arrow) is shown (lane 1) with specific antibody (lane 3; supershift = upper arrow), non-specific (NS) antibody (lane 2), and oligo competitors (WT, lanes 4, 6; MT, lanes 5, 7). **f**, Assays (see **d**) with rDNA reporters containing WT or MT Runx2 sites (A, B, independent clones) and pcDNA vector control showing transcription ratios with or without Runx2. **g**, Binding assay (see **e**) using *in vitro* transcribed and translated (IVTT) recombinant WT or DNA-binding-mutant (DBM) Runx2. **h**, Northern blotting (top) of rDNA-reporter upon co-expression with WT or DBM Runx2 or empty vector (EV) (GAPDH, control). Western blotting (bottom) of epitope-tagged WT and DBM Runx2 proteins using actin as control. Bars denote standard error ($n = 2$) in **a**, **b**, **d**, **f**.

promoter (Fig. 2b). Runx2 and UBF1 binding throughout the *rDNA* repeat are consistent with their co-localization at nucleolar organizing regions and RNA Pol I transcription sites.

We examined the interaction of Runx2 and UBF1 during the cell cycle at multiple sites within the *rDNA* repeat. Occupancy of the proteins increases during G0/G1 and G1/S transitions (Supplementary Fig. 7e–i), but not during mitosis/G1 (Supplementary Fig. 7a–d). Runx2 binding shifts from within the transcriptional initiation region during M/G1 and G1/S into the 5′ transcribed region during G0/G1. The spatial organization of UBF1 across the *rDNA* repeat is more static during changes in rRNA synthesis with preferential binding at enhancer and transcriptional initiation regions (Supplementary Fig. 7j, k). Our results demonstrate that Runx2 binds to the *rDNA* repeat and exhibits site-specific preference within the *rDNA* repeat.

To determine whether Runx2 directly affects *rDNA* transcription, we used a minimal mouse rRNA-promoter reporter[17] that contains a single proximal Runx binding site, as evidenced by electrophoretic mobility shift assays (Fig. 2c, e). Increasing levels of Runx2 cause a dose-dependent repression of the wild-type rDNA promoter (Fig. 2d), but not upon mutation of the Runx element (Fig. 2f). Furthermore, a mutant Runx2 protein that is defective for DNA binding (DBM; Fig. 2g) cannot mediate repression of the rDNA promoter as shown by northern blot analysis (Fig. 2h). These results establish that Runx2 directly regulates *rDNA* gene transcription.

To establish whether Runx2 modulates endogenous rRNA transcription *in vivo*, we examined rRNA synthesis in osteogenic mesenchymal precursors in which Runx2 expression has been genetically ablated. Metabolic labelling experiments reveal enhanced rRNA synthesis in Runx2 null cells, but no significant effect on rRNA processing (Fig. 3a, c). Reverse transcription–quantitative polymerase chain reaction (RT–qPCR) analysis of pre-rRNA (Fig. 3b) and nuclear run-on analysis of RNA transcription (data not shown) further confirm this enhancement in rRNA synthesis in Runx2 null cells. In addition, pulse-labelling of cellular protein with radioactive (^{3}H- or ^{35}S-) methionine is increased in Runx2 null cells, indicating increased capacity for protein synthesis (Fig. 3d, e). Importantly, we find that short interfering RNA (siRNA) mediated knockdown of Runx2 protein levels results in enhanced rRNA expression (Fig. 3f). Taken together, our data indicate that Runx2 functions to inhibit rRNA expression.

Functional effects of Runx2 on the rDNA promoter and colocalization of Runx2 with UBF1 suggest that these proteins interact *in vivo*. Indeed, co-immunoprecipitation assays in several osteoblastic cell lines show that endogenous Runx2 and UBF1 are part of the same complex (Fig. 4a). Notably, Runx2 associates with the transcriptionally active subset of UBF1 that is phosphorylated on Ser 484 (Fig. 4b), as well as with the p48 and p95/p110 subunits of the SL1 complex that is required for promoter recognition by RNA Pol I (Fig. 4b and data not shown). To assess whether Runx2 is recruited to the *rDNA* regulatory region together with UBF1 and the SL1 complex, we performed sequential chromatin immunoprecipitation assays (that is, ChIP-reChIP) (Supplementary Fig. 8). Our results show that UBF1 occupied *rDNA* fragments are simultaneously bound by both Runx2 and p48 (Fig. 4c). Consistent with our microscopic observations (Fig. 1a), the co-occupancy of these proteins at *rDNA* loci also occurs in mitotic cells (Fig. 4c).

Runx proteins are known to regulate gene expression through interactions with chromatin remodelling enzymes. We hypothesized that Runx2 may attenuate rRNA expression by affecting epigenetic modifications at *rDNA* loci. Therefore, we carried out siRNA knockdown experiments to examine whether Runx2 controls post-translational modifications of histones. We performed chromatin immunoprecipitation assays with siRNA treated human Saos cells using antibodies directed against K4-dimethylated histone H3, as well as acetylated histones H3 and H4, all of which are related to gene activation[18,19]. Analysis by qPCR reveals that knockdown of Runx2 protein significantly increases K4-dimethylated histone H3 and acetylated histone H4 near the *rDNA* transcription start-site (human rDNA repeat 1; Fig. 4d). These results support our finding that Runx2 deficiency increases rRNA synthesis (Fig. 3). Hence, Runx2 may attenuate rRNA transcription in part by affecting chromatin modifications. We conclude that Runx2 binds to *rDNA* loci together with UBF1 and components of the essential SL1

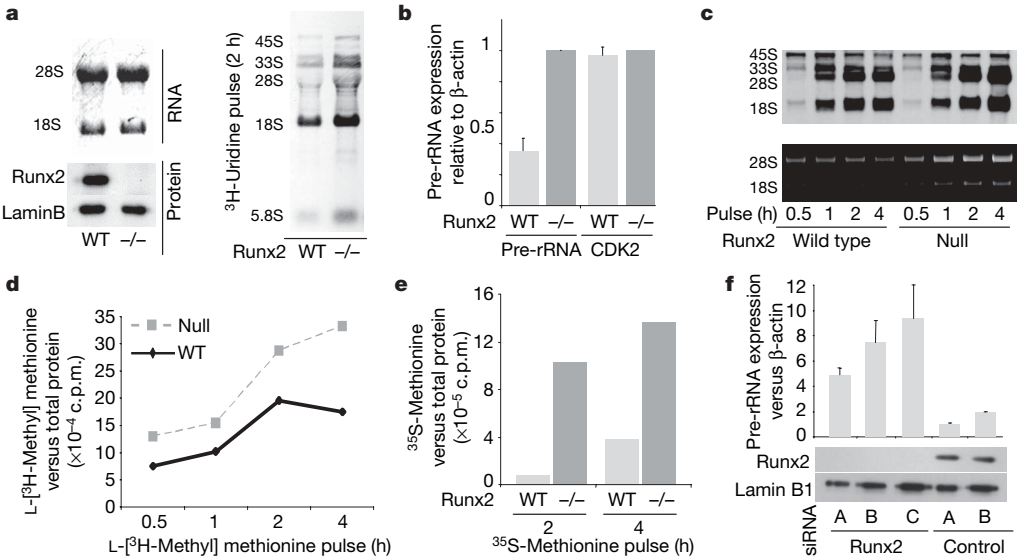


Figure 3 | Runx2 deficiency alters rRNA synthesis *in vivo*. **a**–**e**, Primary calvarial cells from homozygous mouse embryos (17.5 days post-coitus) with WT or null Runx2 alleles. **a**, Panels show rRNA levels by ethidium bromide (top left), Runx2 protein by western analysis (bottom left) and ^{3}H-uridine labelling of RNA synthesis by autoradiography (right). **b**, Pre-rRNA synthesis in equal cell numbers was analysed by RT–qPCR relative to β-actin in WT and null cells (control, Cdk2 levels). **c**, RNA synthesis measured by ^{3}H-methyl-methionine-incorporation at 0.5, 1, 2 and 4 h using autoradiography (top) and rRNA levels by ethidium bromide staining (bottom). **d**, **e**, Incorporation of ^{3}H-methyl-methionine (**d**) or ^{35}S-methionine (**e**) into proteins measured by scintillation counting and normalized to total protein. **f**, Saos cells transfected with three independent Runx2 siRNAs (A, B or C) or siRNA controls (A, GFP; B, chloramphenicol acetyltransferase (CAT)) were examined for unprocessed rRNA (pre-rRNA synthesis) and β-actin by RT–qPCR analysis (top), and Runx2 and LaminB1 protein expression by western blot analysis (bottom). Bars denote standard error ($n = 2$) in **b**, **f**.

complex to provide cell-type specific regulation of rRNA synthesis (Fig. 4e).

We report that Runx2 directly associates with ribosomal DNA loci during interphase and mitosis, and interacts with UBF1 and the SL1 complex to regulate ribosomal RNA synthesis. Retention of Runx2 at nucleolar organizing regions on mitotic chromosomes provides a basis for conveying lineage-specific control of rRNA gene expression to progeny cells. These fundamental findings have major biological and biomedical ramifications, and extend previous studies that implicate cancer-related genes in RNA Pol I transcription[12,17,20–26].

The functional linkage between Runx2 control of rRNA synthesis, proliferation and differentiation provides insight into the tissue-specific phenotype associated with Treacher Collins syndrome. Craniofacial bone defects and growth retardation in Treacher Collins syndrome are linked with deregulated ribosome production[27] and resemble skeletal abnormalities observed in cleidocranial dysplasia, which is caused by Runx2 loss-of-function mutations[28]. The phenotypic penetrance of these diseases can now be interpreted within the context of bone lineage-restricted regulation of ribosomal biogenesis by Runx2.

Our discovery that Runx2 regulates ribosomal biogenesis, which is intricately connected with cell growth, suggests that Runx proteins may establish cell identity by coordinately controlling growth, proliferation and differentiation. Notably, the leukaemia-related Runx1 protein, which is required for haematopoiesis, can also associate with rDNA related nucleolar organizing regions (G.S.S. *et al.*, unpublished observations). Thus, from a broader biological perspective, lineage-specific control of ribosomal biogenesis may be a fundamental function of transcription factors that govern cell fate.

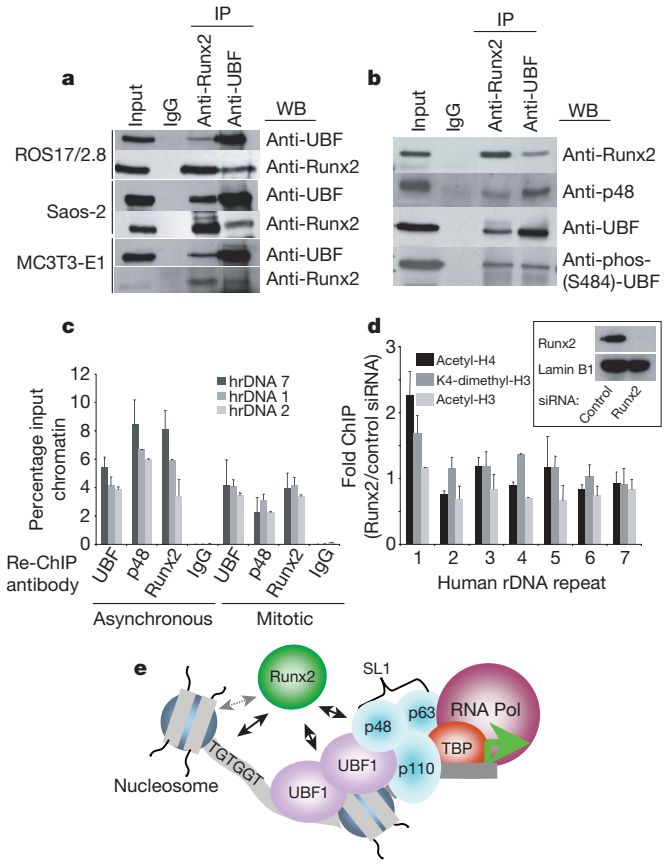


Figure 4 | Runx2 associates with components of the RNA Pol I regulatory complex at *rDNA* loci. Immunoprecipitates (IP) of endogenous Runx2 and UBF1 from osteoblastic cells (**a**) and HA–Runx2 and Flag–UBF1 expressed in NIH-3T3 cells (**b**) react with antibodies against Runx2, UBF1, phospho-UBF1, or the p48 subunit of the SL1 complex in western blots (WB). **c**, ChIP-reChIP assays with proteins endogenous to asynchronous and mitotic Saos cells using UBF1 antibody (primary ChIP) and second immunoprecipitation (reChIP) with antibodies directed against UBF1, p48 and Runx2 or IgG. Chromatin samples were analysed by qPCR using primers flanking the *rDNA* transcription start site, hrDNA-7, hrDNA-1, hrDNA-2 (see Supplementary Table 1 and Supplementary Fig. 6). **d**, ChIPs from asynchronous Saos-2 pre-treated with Runx2 siRNA oligos or non-silencing controls using antibodies against acetylated-histone H4, acetylated-histone H3, and K4-dimethylated-histone H3 were analysed by qPCR and expressed as a ratio. Runx2 knockdown was confirmed by western analysis (inset). **e**, Diagram depicting lineage-specific regulation of rRNA synthesis. Bars denote standard error ($n = 2$) in **c**, **d**.

METHODS

Cell synchronization. Cells were blocked in mitosis by overnight nocodazole treatment followed by shake-off to detach mitotic cells. Where indicated, cells were washed and replated in growth media for mitotic release. Chromosome spreads were generated by brief incubation of mitotic cells in hypotonic KCl buffer followed by centrifugation onto positively charged glass slides. The G0/G1 and G1/S synchronizations were obtained by serum deprivation and re-stimulation of MC3T3 cells. See Supplementary Information for further details.

Microscopy. Cells grown on coverslips as well as chromosome spread preparations were processed for *in situ* immunofluorescence as described[4]. Immunoelectron microscopy was performed essentially as described[29]. See Supplementary Information for further details.

Chromatin immunoprecipitation analysis. Chromatin immunoprecipitation assays (ChIPs) were performed as described[30]. Primers are outlined in Supplementary Table 1. See Supplementary Information for further details.

Co-immunoprecipitation and western blot analysis. Co-immunoprecipitations were performed with UBF (F-9 or H-300, Santa Cruz Biotech), Runx2 (M-70, Santa Cruz Biotech), phospho-specific UBF (Ser 484, Santa Cruz Biotech), or TAFI/p48 (M-19, Santa Cruz Biotech). Western blots were performed using standard techniques. See Supplementary Information for further details.

qPCR analysis of pre-rRNA expression. Total RNA was isolated using Trizol reagent (Invitrogen), column purified, and cDNA was generated in a reverse transcription reaction with random hexamer or gene-specific primers. cDNA was then subjected to real-time PCR using SYBR chemistry with primers (see Supplementary Table 1) flanking early rRNA processing sites.

Metabolic labelling. ^{3}H-Uridine, ^{3}H-methyl-methionine, and ^{35}S-methionine labelling experiments were carried out essentially as described[24]. See Supplementary Information for further details.

siRNA and plasmid DNA transfection. siRNA transfections were performed using standard techniques with the following siRNA duplexes: Runx2-A r(GGU UCA ACG AUC UGA GAU U)d(TT), Runx2-B r(UCU GUU UGG CGA CCA UAU U)d(TT), Runx2-C r(UGC CUC UGC UGU UUG AAA)d (TT). MR170-BH rDNA-promoter reporter plasmid is described elsewhere[17]. MR170-BH reporter with a mutant Runx site was generated by site-directed mutagenesis using the following primer (5′-GTT GTT CCT TTG A*AC* TCC GGT TCT TT). Reporter expression was monitored by qPCR (see Supplementary Table 1 for primers) and by northern blot analysis using the pUC9 DNA reporter as a hybridization probe. See Supplementary Information for further details.

Electrophoretic mobility shift assays. See Supplementary Information for details.

Received 9 August; accepted 20 November 2006.

1. Russell, J. & Zomerdijk, J. C. RNA-polymerase-I-directed rDNA transcription, life and works. *Trends Biochem. Sci.* **30**, 87–96 (2005).
2. Blyth, K., Cameron, E. R. & Neil, J. C. The runx genes: gain or loss of function in cancer. *Nature Rev. Cancer* **5**, 376–387 (2005).
3. Zaidi, S. K. *et al.* Mitotic partitioning and selective reorganization of tissue specific transcription factors in progeny cells. *Proc. Natl Acad. Sci. USA* **100**, 14852–14857 (2003).
4. Zaidi, S. K. *et al.* A specific targeting signal directs Runx2/Cbfa1 to subnuclear domains and contributes to transactivation of the osteocalcin gene. *J. Cell Sci.* **114**, 3093–3102 (2001).
5. Javed, A. *et al.* Groucho/TLE/R-Esp proteins associate with the nuclear matrix and repress RUNX (CBFα/AML/PEBP2α) dependent activation of tissue-specific gene transcription. *J. Cell Sci.* **113**, 2221–2231 (2000).
6. Martinez-Balbas, M. A. *et al.* Displacement of sequence-specific transcription factors from mitotic chromatin. *Cell* **83**, 29–38 (1995).
7. Muchardt, C. *et al.* The hbrm and BRG-1 proteins, components of the human SNF/SWI complex, are phosphorylated and excluded from the condensed chromosomes during mitosis. *EMBO J.* **15**, 3394–3402 (1996).

8. Nuthall, H. N., Joachim, K., Palaparti, A. & Stifani, S. A role for cell cycle-regulated phosphorylation in Groucho-mediated transcriptional repression. *J. Biol. Chem.* **277,** 51049–51057 (2002).
9. Prasanth, K. V., Sacco-Bubulya, P. A., Prasanth, S. G. & Spector, D. L. Sequential entry of components of gene expression machinery into daughter nuclei. *Mol. Biol. Cell* **14,** 1043–1057 (2003).
10. Gonzalez, I. L. & Sylvester, J. E. Human rDNA: evolutionary patterns within the genes and tandem arrays derived from multiple chromosomes. *Genomics* **73,** 255–263 (2001).
11. Galindo, M. *et al.* The bone-specific expression of RUNX2 oscillates during the cell cycle to support a G1 related anti-proliferative function in osteoblasts. *J. Biol. Chem.* **280,** 20274–20285 (2005).
12. Javed, A. *et al.* Impaired intranuclear trafficking of Runx2 (AML3/CBFA1) transcription factors in breast cancer cells inhibits osteolysis *in vivo*. *Proc. Natl Acad. Sci. USA* **102,** 1454–1459 (2005).
13. Mais, C. *et al.* UBF-binding site arrays form pseudo-NORs and sequester the RNA polymerase I transcription machinery. *Genes Dev.* **19,** 50–64 (2005).
14. Gebrane-Younes, J., Fomproix, N. & Hernandez-Verdun, D. When rDNA transcription is arrested during mitosis, UBF is still associated with non-condensed rDNA. *J. Cell Sci.* **110,** 2429–2440 (1997).
15. Roussel, P. *et al.* Localization of the RNA polymerase I transcription factor hUBF during the cell cycle. *J. Cell Sci.* **104,** 327–337 (1993).
16. Cheutin, T. *et al.* Three-dimensional organization of active rRNA genes within the nucleolus. *J. Cell Sci.* **115,** 3297–3307 (2002).
17. Budde, A. & Grummt, I. p53 represses ribosomal gene transcription. *Oncogene* **18,** 1119–1124 (1999).
18. Jenuwein, T. & Allis, C. D. Translating the histone code. *Science* **293,** 1074–1080 (2001).
19. Martin, C. & Zhang, Y. The diverse functions of histone lysine methylation. *Nature Rev. Mol. Cell Biol.* **6,** 838–849 (2005).
20. Hannan, K. M. *et al.* Rb and p130 regulate RNA polymerase I transcription: Rb disrupts the interaction between UBF and SL-1. *Oncogene* **19,** 4988–4999 (2000).
21. Hannan, K. M. *et al.* RNA polymerase I transcription in confluent cells: Rb downregulates rDNA transcription during confluence-induced cell cycle arrest. *Oncogene* **19,** 3487–3497 (2000).
22. Voit, R., Schafer, K. & Grummt, I. Mechanism of repression of RNA polymerase I transcription by the retinoblastoma protein. *Mol. Cell. Biol.* **17,** 4230–4237 (1997).
23. Cavanaugh, A. H. *et al.* Activity of RNA polymerase I transcription factor UBF blocked by Rb gene product. *Nature* **374,** 177–180 (1995).
24. Grandori, C. *et al.* c-Myc binds to human ribosomal DNA and stimulates transcription of rRNA genes by RNA polymerase I. *Nature Cell Biol.* **7,** 311–318 (2005).
25. Arabi, A. *et al.* c-Myc associates with ribosomal DNA and activates RNA polymerase I transcription. *Nature Cell Biol.* **7,** 303–310 (2005).
26. Poortinga, G. *et al.* MAD1 and c-MYC regulate UBF and rDNA transcription during granulocyte differentiation. *EMBO J.* **23,** 3325–3335 (2004).
27. Valdez, B. C. *et al.* The Treacher Collins syndrome (TCOF1) gene product is involved in ribosomal DNA gene transcription by interacting with upstream binding factor. *Proc. Natl Acad. Sci. USA* **101,** 10709–10714 (2004).
28. Otto, F., Kanegane, H. & Mundlos, S. Mutations in the *RUNX2* gene in patients with cleidocranial dysplasia. *Hum. Mutat.* **19,** 209–216 (2002).
29. Nickerson, J. A., He, D. C., Krochmalnic, G. & Penman, S. Immunolocalization in three dimensions: immunogold staining of cytoskeletal and nuclear matrix proteins in resinless electron microscopy sections. *Proc. Natl Acad. Sci. USA* **87,** 2259–2263 (1990).
30. Hovhannisyan, H. *et al.* Maintenance of open chromatin and selective genomic occupancy at the cell-cycle-regulated histone H4 promoter during differentiation of HL-60 promyelocytic leukemia cells. *Mol. Cell. Biol.* **23,** 1460–1469 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank I. Grummt for rDNA reagents and J. Rask for editorial assistance. We also thank A. Pardee for discussions. Studies reported were in part supported by the National Institutes of Health.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to G.S.S. (gary.stein@umassmed.edu).

LETTERS

The APOBEC-2 crystal structure and functional implications for the deaminase AID

Courtney Prochnow[1]*, Ronda Bransteitter[1]*, Michael G. Klein[1], Myron F. Goodman[1] & Xiaojiang S. Chen[1]

APOBEC-2 (APO2) belongs to the family of apolipoprotein B messenger RNA-editing enzyme catalytic (APOBEC) polypeptides, which deaminates mRNA and single-stranded DNA[1,2]. Different APOBEC members use the same deamination activity to achieve diverse human biological functions. Deamination by an APOBEC protein called activation-induced cytidine deaminase (AID) is critical for generating high-affinity antibodies[3], and deamination by APOBEC-3 proteins can inhibit retrotransposons and the replication of retroviruses such as human immunodeficiency virus and hepatitis B virus[4–7]. Here we report the crystal structure of APO2. APO2 forms a rod-shaped tetramer that differs markedly from the square-shaped tetramer of the free nucleotide cytidine deaminase, with which APOBEC proteins share considerable sequence homology. In APO2, two long α-helices of a monomer structure prevent the formation of a square-shaped tetramer and facilitate formation of the rod-shaped tetramer via head-to-head interactions of two APO2 dimers. Extensive sequence homology among APOBEC family members allows us to test APO2 structure-based predictions using AID. We show that AID deamination activity is impaired by mutations predicted to interfere with oligomerization and substrate access. The structure suggests how mutations in patients with hyper-IgM-2 syndrome inactivate AID, resulting in defective antibody maturation.

We crystallized APO2, which contains amino acid residues 41–224, with four monomers in each asymmetric unit that form a tetramer with an atypical elongated shape (Fig. 1a). This tetramer

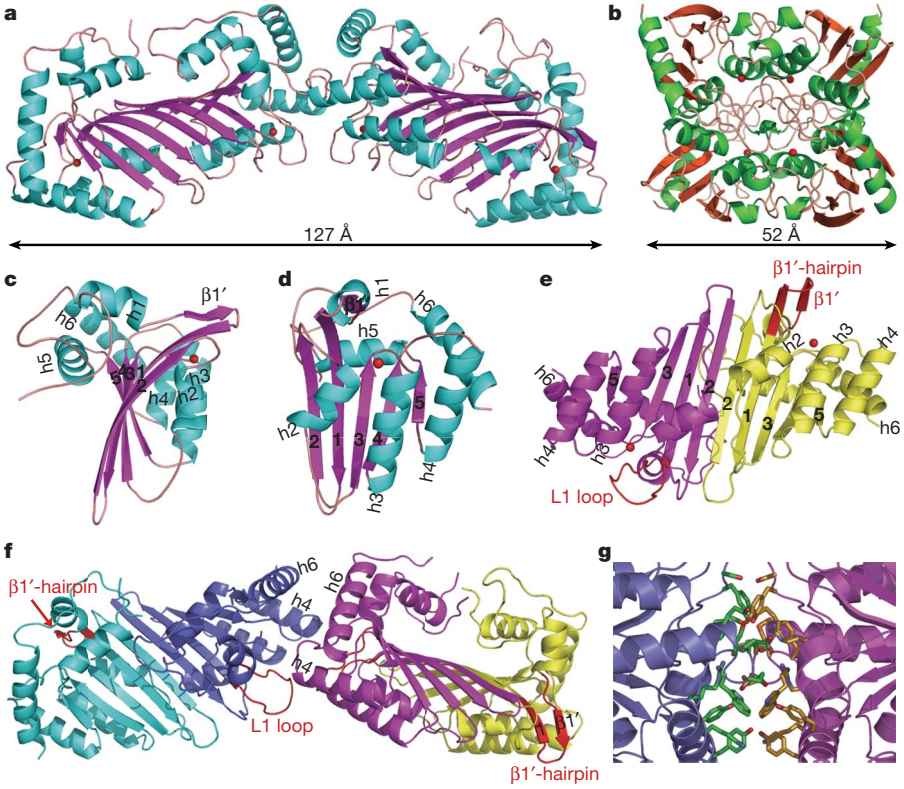


Figure 1 | The structure of APO2. a, The APO2 tetramer structure. It has an end-to-end span of ~126.9 Å. Zn atoms in the active centres are shown as red spheres. **b**, The square-shaped structure of human cytidine deaminase (PDB accession number: 1MQ0), a fntCDA. **c**, **d**, The APO2 monomer structure rotated by 90 degrees, showing the unique features of APO2: the short β1′ strand and helices h4 and h6. h4 and h6 dictate how APO2 oligomerizes. **e**, The APO2 dimer formed by two monomers (in purple and yellow). Each has a different conformation for the h1/β1-turn (in red): a loop (L1) and a hairpin. **f**, The tetrameric interface, showing the extensive interactions mediated through h4, h6 and L1. **g**, The stick model (hydrophobic, polar and charged amino acids in h4, h6 and L1) of the interactions at the tetramer interface.

[1]Molecular and Computational Biology, University of Southern California Los Angeles, California 90089, USA.
*These authors contributed equally to this work.

assembles through two different monomer–monomer interfaces, in contrast to the canonical square shape of the free nucleotide cytidine deaminase (fntCDA) tetramer (Fig. 1b), in which all four monomers interact with each other[8]. The elongated APO2 tetramer has the shape of a butterfly (Fig. 1a) with an end-to-end span of approximately 126.9 Å.

The APO2 monomer appears to adopt the typical core fold of the fntCDAs with a five-stranded β-sheet flanked by helices on both sides (Figs 1c, d). However, one new attribute is the additional α-helices surrounding the core β-sheet (Figs 1c, d); six long helices are present in the APO2 monomer whereas only three or four are observed in the fntCDA monomer (excluding the shorter 3_{10} helices)[8–11]. Helices h3 and h6 make extensive contacts with h4, stabilizing the position of the helices within the monomer subunit. On the basis of the close sequence homology of APO2 with other APOBEC proteins, the long helix (h4) probably serves as a structural signature of this family (Figs 1c, d).

The APO2 dimer is formed by pairing two long β-strands (β2) (Fig. 1e), joining two β-sheets sideways to form one wide β-sheet that resembles a ribcage (Fig. 1a, e). Twelve residues (residues 82–93) on each β2 strand form 12 hydrogen bonds through main-chain atoms, providing the principal bonding force between the two monomers. The dimer interface is reinforced by the side-chain interactions occurring through the loops and helices located on both sides of the β-sheet. Ordered water molecules also help to stabilize this interface.

The dimer is nearly symmetrical (Fig. 1e) with six helices (h2, h3 and h4 of both molecules) located on one side of the augmented β-sheet and four helices (h1 and h5 of both monomers) on the other side. Capped on both edges of the β-sheet are h4 and h6. However, one part of the dimer shows obvious asymmetry at the turn between h1 and strand β1 (h1/β1-turn). This h1/β1-turn (residues 57–68) assumes a hairpin structure (β1′-hairpin) in one monomer, and a loop conformation (L1) in the other monomer (Fig. 1e, f).

The APO2 tetramer is formed by two dimers joining through head-to-head interactions. The two dimers make extensive contacts via the residues from h4 and h6, as well as the loop L1 at the h1/β1-turn (Fig. 1f). Residues Y61, F155, M156, W157, P160, Y214 and Y215 from each side of the interface form extensive hydrophobic packing interactions, and residues R57, S62, S63, R153, E158, E159 and E161 establish salt bridges and hydrogen bonds (Fig. 1g). Some charged residues even use their aliphatic side chains to interact with hydrophobic residues. Thus, hydrophobic, polar and charged amino acid side chains are all involved in the tetramerization interactions. The total buried area is 1,745 $Å^2$ within the tetramer interface, where h4 and h6 play a major role forming the interface (Fig. 1f). h4 and h6 also sterically hinder the formation of the square-shaped fntCDA-type tetramer by occupying the space where another monomer would need to be. Therefore, h4 and h6 appear to determine directly the elongated tetramer formation.

A prominent feature of the APO2 tetramer distinctive from the fntCDA tetramer is that the active sites are accessible for large RNA or

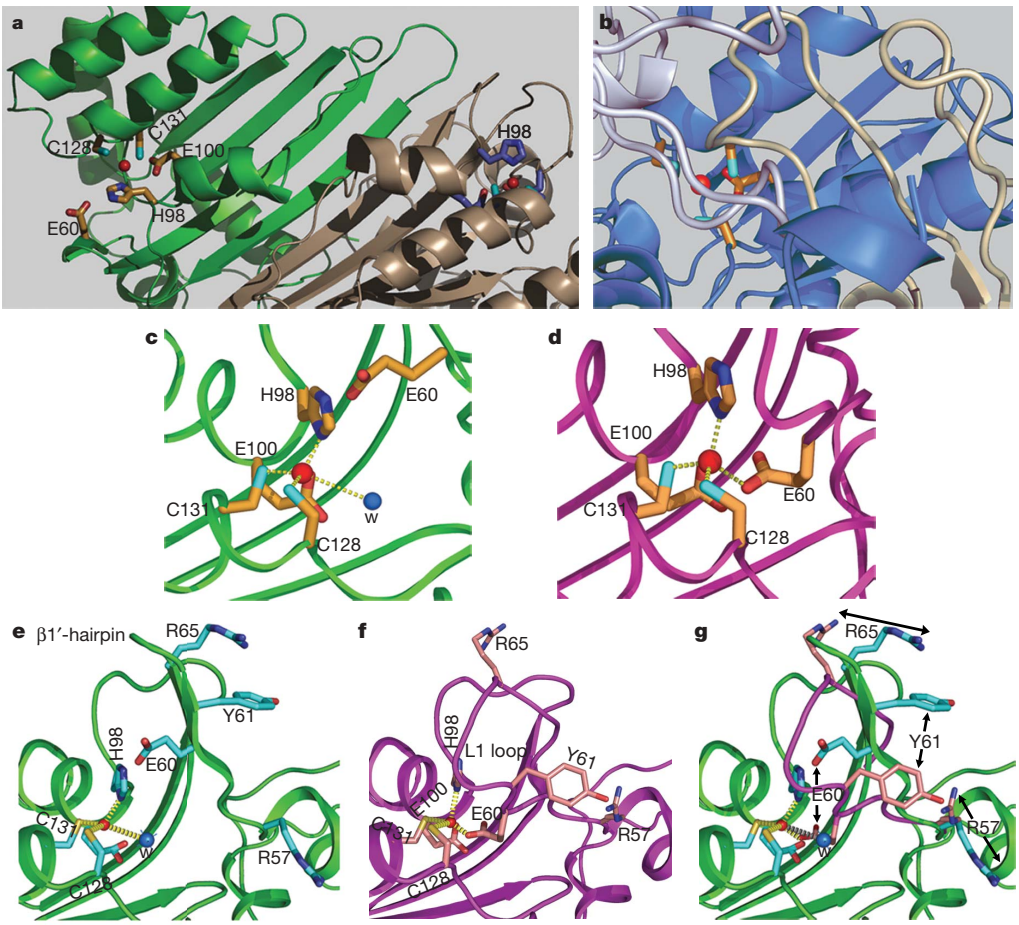


Figure 2 | The APO2 active site. **a**, The APO2 active sites are accessible to DNA/RNA. Red spheres represent Zn. **b**, The fntCDA active site is accessible only to free nucleotides. **c**, The outer APO2 active sites show Zn coordination (yellow dashed lines) by three residues (H98, C128, C131) and a water molecule (blue sphere). **d**, The middle APO2 active centre sites show Zn coordination by a fourth residue, E60. **e**, In the β1′-hairpin structure, the hydrophobic ring of Y61 interacts with the guanidine group of R65, stabilizing the conformation. **f**, In the h1/β1 loop, the E60 coordinates with Zn. Y61 now rotates away from R65 and interacts with R57, facilitating the disruption of the β1′-hairpin and stabilizing the loop conformation. **g**, Superimposed monomers show that the h1/β1 loop (purple) is pulled down ~8.5 Å towards the active site owing to the E60–Zn bond formation.

DNA substrates (Fig. 2a). In the square-shaped fntCDA tetramer, loops from two neighbouring monomers cover the active sites so that only small free nucleotides can bind to the buried sites (Fig. 2b). Although the yeast fntCDA, CDD1, has been reported to deaminate the apolipoprotein B mRNA *in vitro*, its known biological substrate *in vivo* is a free nucleotide and the CDD1 structure is a canonical square-shaped tetramer[9].

In fntCDAs, the active centre Zn atom is coordinated by three residues (either three cysteines, or two cysteines + one histidine) and forms a fourth bond with a water molecule with a bond distance of ~3.0 Å (ref. 10). This type of Zn coordination is also present in APO2 (Fig. 2c), but only in the two outer monomers of the tetramer. Surprisingly, the active sites for the other two monomers in the middle of a tetramer contain an E60 residue, which replaces the water molecule and makes the fourth coordination bond with the Zn (Fig. 2d). This coordination of Zn by four residues is unexpected given that all known fntCDA structures have only three amino acid residues participating in Zn coordination[8–13].

A closer examination of the structure reveals a 'built-in' mechanism for a conformational switch between the two types of Zn coordination. The switch is mediated by sequences contained in the h1/β1-turn, in which E60 is located. The h1/β1-turn can adopt either a hairpin (β1′-hairpin, Fig. 2e) or a loop (L1) conformation (Fig. 2f), which controls whether or not E60 coordinates with Zn. The E60 is located 6 Å from the Zn when the h1/β1-turn is a β1′-hairpin (Fig. 2e). The β1′-hairpin is stabilized by main-chain hydrogen bonds within the β1′-hairpin and reinforced by interactions between Y61 and the guanidine group of R65. In the two middle monomers of an APO2 tetramer, the h1/β1-turn folds into a loop (L1, Fig. 1f). In this conformation, Y61 rotates its side chain to interact with R57 instead of the R65 (Fig. 2f, g). The new pairing of Y61 with R57 destabilizes the β1′-hairpin while stabilizing the loop. In the loop conformation, the E60 is 2.2 Å from the Zn (Fig. 2f).

The hairpin–loop switch may have two important consequences. First, switching to the loop and forming the fourth Zn coordination by E60 prevents coordination by water and subsequent hydroxylation of Zn necessary for deamination (Fig. 2d, f). Second, Zn coordination by E60 pulls the h1/β1-turn approximately 8.5 Å towards the active centre (Fig. 2g), which could restrict substrate access to the active centre. On the other hand, breaking of the fourth coordination of E60 may allow the loop to move away from the active centre to form the β1′-hairpin as observed in the outer monomers. The E60 would no longer prevent the Zn hydroxylation and nucleic acid substrate access to those active sites. Thus, the hairpin–loop switch can be a molecular mechanism for regulating substrate access and enzyme activity mediated through Zn coordination.

The APO2 fragment in the structure shares a 33.3% amino acid identity (44.6% homology) with AID, and the buried residues in APO2 share a 75% identity (96% homology) with AID (Fig. 3a). The highly conserved residues buried inside the structure and those located at the dimer/tetramer interfaces strongly suggest a structural conservation of AID with APO2. Thus, the APO2 crystal structure should provide functional insights for AID and other APOBEC family members, despite of the lack of the known biological activity of APO2. For this reason, we use AID as a surrogate to test how mutations guided by APO2 structure affect AID deamination activity.

We generated glutathione *S*-transferase (GST)–AID mutants with amino acid substitutions located at the tetrameric interface (Fig. 3b), and showed that the mutants either had no detectable or significantly reduced deaminase activity compared to wild-type GST–AID (Fig. 3d–f). Mutants R112C and Y114A/F115A were inactive (Fig. 3e, f), while mutants K16A and C116A had a 3.3-fold reduction in activity (Figs 3e, f).

AID mutations within the predicted dimerization domain, F46A/Y48A (Fig. 3c), displayed a fourfold decrease in deamination activity

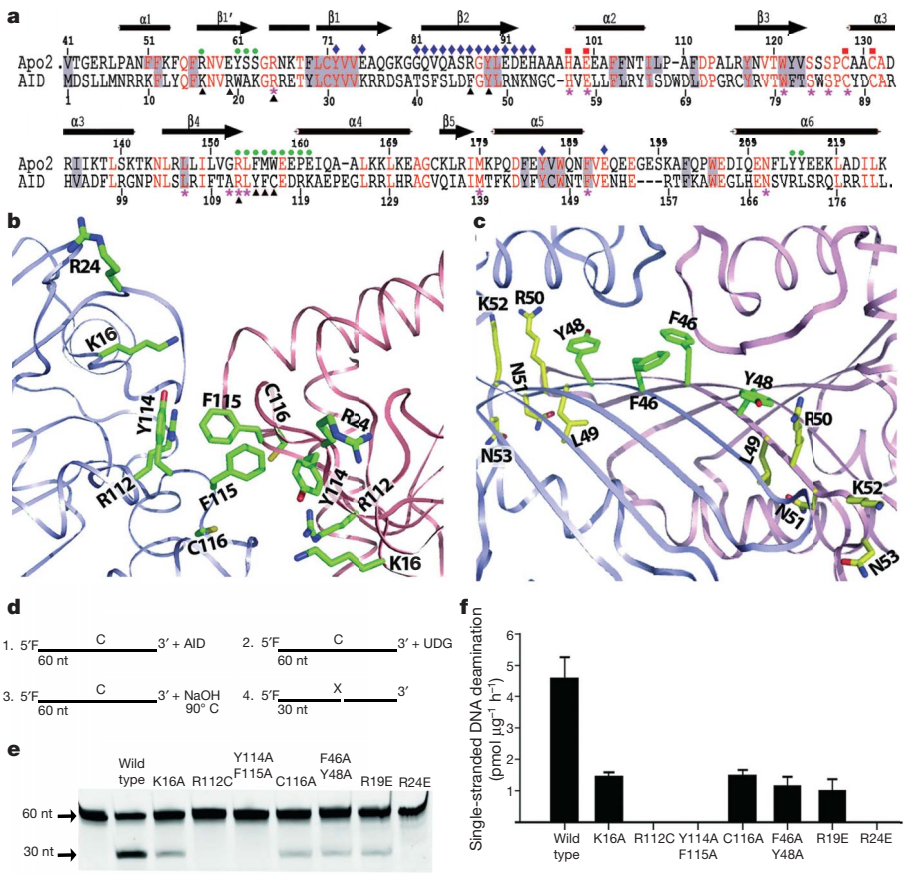


Figure 3 | Structurally guided mutagenesis of AID impairs deamination activity. **a**, A sequence alignment of APO2 and AID, showing significant homology. Red, identical residues; grey shading, buried residues; red squares, active centre residues; green dots, tetrameric interface residues; blue diamonds, dimeric interface residues; purple stars, HIGM mutated residues; and black triangles, mutated AID residues. **b**, Mutated AID residues (in green) at the tetramer interface as modelled based on the APO2 structure. **c**, Mutated AID residues (in green) in the dimer interface as modelled based on the APO2 structure. **d**, A sketch describing the cytidine deamination assay. F, fluorescein; UDG, is uracil DNA glycosylase. **e**, Denaturing PAGE analysis of the deamination activity for wild-type and mutant AID proteins. The 30-nucleotide (nt) band indicates deamination activity. **f**, Bar representation of the specific activities for wild-type and mutant AID proteins.

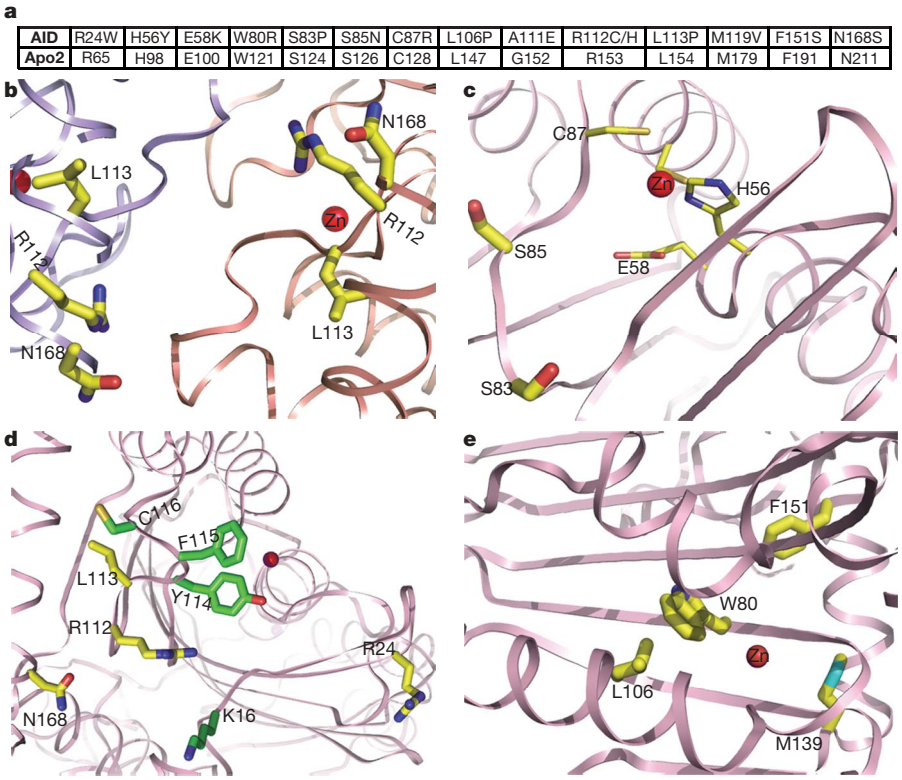


Figure 4 | AID HIGM-2 mutations. a, Alignment of mutated residues of AID from HIGM-2 patients with the corresponding residues in APO2, showing high sequence conservation. **b**, Mapping the residues in AID HIGM-2 mutations (R112, L113, N168) to the tetramer interface as modelled from the APO2 structure. **c**, Mapping the AID HIGM-2 mutations, S83 and S85, near the active site. **d**, Mapping the AID mutations, K16, Y114/F115 and C116 (in green), to the exposed surface of an outer monomer. The HIGM-2 AID residues (R112, L113, N168, in yellow), which are at the tetramer interface (**b**), are also located on this exposed surface. **e**, Mapping of AID HIGM-2 mutations, W80, L106, M139 and F151, to the interior core structure.

(Fig. 3e, f). The dimer interface is extensive, so two mutations should not completely disrupt dimeric AID. This explains why weak deamination activity was observed with this double AID mutant. These mutational results suggest that the residues within the predicted dimeric and tetrameric interfaces are important for deamination activity. One caveat is that the residues on the tetramer interface are also present on the exposed surface of the outer ends of the tetramer and thus could also be involved in an additional role beside tetramerization (Figs 3b, 4d). We noticed that in gel filtration assays the tetramer was a minor species when compared with the dimer, suggesting a stronger dimeric interaction.

AID has an arginine (R19) at the equivalent position of the APO2 E60 residue (Fig 3a) that may have a negative regulatory role for APO2 activity by blocking Zn hydroxylation and substrate access (Fig. 2f). We showed that an AID R19E mutant mimicking APO2 E60 had a significantly decreased deamination activity (about 4.6-fold less than the wild type, Figs 3e, f). Similarly, the AID R24 residue is equivalent to APO2 R65, which interacts with Y61 of APO2 to stabilize the open β1′-hairpin conformation. We predicted that the disruption of the R65–Y61 interaction would collapse the β1′-hairpin into the closed loop conformation to block substrate access and impair deamination activity. Indeed, the AID R24E mutant was completely inactive on single-stranded DNA (Fig. 3e, f).

Mutations in human AID cause hyper-IgM-2 (HIGM-2) syndrome, characterized by an impaired production of high-affinity antibodies[14,15]. The mutated AID residues of HIGM-2 patients are highly conserved in APO2 (Fig. 4a). A plausible explanation for why and how HIGM-2 mutations disrupt AID function is given by the structure of APO2 (Fig. 4b–e). On the basis of the crystal structure, HIGM-2 AID mutations can be divided into four classes. The first mutant class (A111, R112, L113 and N168) occurs at the tetramerization interface (Fig. 4b). The second mutant class includes residues in and near the active centre (Fig. 4c), H56, E58, S83, S85 and C87, which are conserved among all APOBEC enzymes. The AID R24 residue is also mutated in HIGM-2 patients. As previously discussed, R24 may stabilize the β1′-hairpin, which keeps the active site open for DNA/RNA access. The third class consists of residues located on the enzyme surface (Fig. 4d), including those residues located at the tetramer interface (A111, R112, L113, N168; see Fig. 4d). A fourth class of HIGM-2 AID mutations are those with large hydrophobic side chains buried within the core (Fig. 4e), including W80, L106, M139 and F151. Three of these residues are located near the active centre. Mutating these residues should disrupt the folding and stability of AID.

Since many of the APOBEC enzymes are reported to form dimers and multimers, the APO2 structure may shed light on how these enzymes oligomerize[16–23]. The elucidation of the APO2 structure, fortified by the structure-guided predictions for the activity of specific AID mutants, provides a structural basis to pursue further functional studies of APOBEC proteins with an eye towards developing therapeutic strategies to deal with deficiency in deaminating cytidine and to restrict retroviral replication.

METHODS

Crystallography statistics can be found in the Supplementary Information.

Protein purification and crystallization. Human APO2 containing residues 41–224 was cloned and expressed in *Escherichia coli* as a recombinant GST fusion protein. Following GST cleavage by thrombin, further purification of APO2 was achieved using Superdex-75 gel filtration chromatography. Native and selenium-methionine labelled protein was concentrated to 15 mg ml^{-1} in a buffer containing 25 mM Hepes, pH 7.0, 50 mM NaCl and 10 mM dithiothreitol. Crystals were grown at 18 °C by hanging-drop vapour diffusion from a reservoir

solution of 85 mM Na-citrate, pH 5.6, 160 mM $LiSO_4$, 24% (weight/volume) polyethylene glycol monomethyl ether and 15% glycerol.

Structure determination and refinement. Native and selenium-multiwavelength anomalous diffraction (MAD) data were collected at the synchrotron and processed using HKL2000[24] (Supplementary Table 1). An initial solution was obtained at 3.5 Å using the program Solve[25] and located eight selenium atoms using the peak wavelength data set. A search with SHELXD[26] found four additional selenium atoms (totalling twelve) and, subsequently, the program SHARP identified four additional weaker anomalous scattering atoms, which were recognized as Zn atoms. Density modification schemes of solvent flattening and fourfold non-crystallographic symmetry (NCS) averaging were applied using the program RESOLVE[27]. Additionally, phase extension in RESOLVE was performed with the native data set to 2.5 Å resolution using the two-wavelength MAD phases calculated in SHARP. The molecular model was built based on this experimental map using the program "O" and was refined with the Crystallography and NMR System (CNS). A twofold NCS constrain was applied during the initial simulated annealing, but the final refinement was carried out without NCS constrain. The protein geometry is excellent when examined using the program PROCHECK.

Construction of AID mutants. Mutant AID proteins were contructed by site-directed mutagenesis using the pGEX-KG-AID vector as the PCR template and primers specific for the respective mutations (5′-CTG AGG ATC TTC ACC GCG TGC CTC TAC TTC TGT GAG GAC-3′ (R112C), 5′-ATC TTC ACC GCG CGC CTC GCC GCC TGT GAG GAC CGC AAG GCT-3′ (Y114/Y115), 5′-GCG CGC CTC TAC TTC TGT GCG GCC CGC AAG GCT GAG CCC GAG-3′ (E117/E118A), 5′-AAG TTT CTT TAC CAA TTC GCA AAT GTC CGC TGG GCT AAG-3′ (K16A), 5′-ACC GCG CGC CTC TAC TTC GCT GAG GAC CGC AAG GCT GAG-3′ (C116A), 5′-TAC CAA TTC AAA AAT GTC GAG TGG GCT AAG GGT CGG CGT-3′ (R19E), and 5′-ACA TCC TTT TCA CTG GAC GCT GGT GCT CTT CGC AAT AAG AAC GGC-3′ (F46A/Y48A). Mutant constructs were verified by DNA sequencing.

Deamination reactions. Deamination experiments were performed as previously described[28] with the exception that the substrate used was a fluorescein-dT incorporated single-stranded DNA substrate (5′-taa agg fluorescein-dTga aga gag gag aga gaa gta agC tga aga gag aga agg aag aga gtg aag gag-3′). Reaction products were visualized on a BioRad FX scanner.

Received 10 October; accepted 28 November 2006.
Published online 24 December 2006.

1. Pham, P., Bransteitter, R. & Goodman, M. F. Reward versus risk: DNA cytidine deaminases triggering immunity and disease. *Biochemistry* **44,** 2703–2715 (2005).
2. Conticello, S. G., Thomas, C. J. F., Petersen-Mahrt, S. K. & Neuberger, M. S. Evolution of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases. *Mol. Biol. Evol.* **22,** 367–377 (2004).
3. Bransteitter, R., Sneeden, J. L., Allen, S., Pham, P. & Goodman, O. M. F. First AID (activation-induced cytidine deaminase) is needed to produce high affinity isotype-switched antibodies. *J. Biol. Chem.* **281,** 16833–16836 (2006).
4. Chiu, Y. L. & Greene, W. C. Multifaceted antiviral actions of APOBEC3 cytidine deaminases. *Trends Immunol.* **27,** 291–297 (2006).
5. Cullen, B. R. Role and mechanism of action of the APOBEC3 family of antiretroviral resistance factors. *J. Virol.* **80,** 1067–1076 (2006).
6. Franca, R., Spadari, S. & Maga, G. APOBEC deaminases as cellular antiviral factors: a novel natural host defense mechanism. *Med. Sci. Monit.* **12,** RA92–RA98 (2006).
7. Bonvin, M. *et al.* Interferon-inducible expression of APOBEC3 editing enzymes in human hepatocytes and inhibition of hepatitis B virus replication. *Hepatology* **43,** 1364–1374 (2006).
8. Johansson, E., Mejlhede, N., Neuhard, J. & Larsen, S. Crystal structure of the tetrameric cytidine deaminase from *Bacillus subtilis* at 2.0 Å resolution. *Biochemistry* **41,** 2563–2570 (2002).
9. Xie, K. *et al.* The structure of a yeast RNA-editing deaminase provides insight into the fold and function of activation-induced deaminase and APOBEC-1. *Proc. Natl Acad. Sci. USA* **101,** 8114–8119 (2004).
10. Teh, A. *et al.* The 1.48 Å resolution crystal structure of the homotetrameric cytidine deaminase from mouse. *Biochemistry* **45,** 7825–7833 (2006).
11. Chung, S. J., Fromme, J. C. & Verdine, G. L. Structure of human cytidine deaminase bound to a potent inhibitor. *J. Med. Chem.* **48,** 658–660 (2005).
12. Betts, L., Xiang, S., Short, S. A., Wolfenden, R. & Carter, C. W. Cytidine deaminase. The 2.3 Å crystal structure of an enzyme: transition-state analog complex. *Curr. Biol.* **235,** 635–656 (1994).
13. Smith, A. A., Carlow, D. C., Wolfenden, R. & Short, S. A. Mutations affecting transition-state stabilization by residues coordinating zinc at the active site of cytidine deaminase. *Biochemistry* **33,** 6468–6474 (1994).
14. Durandy, A., Peron, S. & Fischer, A. Hyper-IgM syndromes. *Curr. Opin. Rheumatol.* **18,** 369–376 (2006).
15. Minegishi, Y. *et al.* Mutations in activation-induced cytidine deaminase in patients with hyper IgM syndrome. *Clin. Immunol.* **97,** 203–210 (2000).
16. Anant, S. *et al.* ARCD-1, an apobec-1-related cytidine deaminase, exerts a dominant negative effect on C to U RNA editing. *Am. J. Cell Physiol.* **281,** C1904–C1916 (2001).
17. Jarmuz, A. *et al.* An anthropoid-specific locus of orphan C to U RNA-editing enzymes on chromosome 22. *Genomics* **79,** 285–296 (2002).
18. Shindo, K. *et al.* The enzymatic activity of CEM15/Apobec-3G is essential for the regulation of the infectivity of HIV-1 virion but not a sole determinant of its antiviral activity. *J. Biol. Chem.* **278,** 44412–44416 (2003).
19. Wiegand, H. L., Doehle, B. P., Bogerd, H. P. & Cullen, B. R. A second human antiretroviral factor, APOBEC3F, is suppressed by the HIV-1 and HIV-2 Vif proteins. *EMBO J.* **23,** 2451–2458 (2004).
20. Opi, S. *et al.* Monomeric APOBEC3G is catalytically active and has antiviral activity. *J. Virol.* **80,** 4673–4682 (2006).
21. Navarro, F. *et al.* Complementary function of the two catalytic domains of APOBEC3G. *Virology* **333,** 374–386 (2005).
22. Wang, J. *et al.* Identification of a specific domain required for dimerization of activation-induced cytidine deaminase. *J. Biol. Chem.* (in the press).
23. Teng, B. *et al.* Mutational analysis of apolipoprotein B mRNA editing enzyme (APOBEC1): structure-function relationships of RNA editing and dimerization. *J. Lipid Res.* **40,** 623–635 (1999).
24. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
25. Terwilliger, T. C. & Berendzen, J. Automated MAD and MIR structure solution. *Acta Crystallogr.* **55,** 849–861 (1999).
26. Schneider, T. R. & Sheldrick, G. M. Substructure solution with SHELXD. *Acta Crystallogr.* **58,** 1772–1779 (2002).
27. Terwilliger, T. C. Maximum-likelihood density modification. *Acta Crystallogr.* **56,** 965–972 (2000).
28. Bransteitter, R., Pham, P., Scharff, M. D. & Goodman, M. F. Activation-induced cytidine deaminase deaminates deoxycytidine on single-stranded DNA but requires the action of RNase. *Proc. Natl Acad. Sci. USA* **100,** 4102–4107 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank L. Chen for comments on the manuscript. We also thank G. Wang from Chen lab and the staff at ALS LBL8.2.1, BL8.2.2 and APS 19id in the Argonne National Laboratory for assistance in data collection. This work was supported in part by National Institutes of Health grants to M.F.G. and X.S.C. and an NIH-NIA Predoctoral Traineeship to R.B.

Author Information The structure of APO2 has been uploaded to the Protein Data Bank under accession number 2NYT and to the Research Collaboratory for Structural Bioinformatics (RCSB) under accession number RCSB040471. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to X.S.C. (xiaojiang.chen@usc.edu).

naturejobs

THE CAREERS MAGAZINE FOR SCIENTISTS

The Association of American Medical Colleges last month approved a document that, if it becomes more widely adopted, could revolutionize postdoctoral training. The association's Compact Between Postdoctoral Appointees and Their Mentors (www.aamc.org/postdoccompact) is powerful because it gives both parties clear responsibilities and sets clear expectations for each side.

It tells postdocs to take primary responsibility for their career development. It states that their research project should be developed with their mentor, and that this should have clearly defined goals and timelines, all of which, ideally, should be agreed at the time the postdoc appointment is made. It asks postdocs to follow good research practices, to adhere to ethical standards and to treat their colleagues with respect. It also challenges them to assume more responsibility as their project progresses and puts the onus on them to request formal performance reviews. And, finally, it charges them to seek professional development activities outside the lab, both in terms of scientific and career development.

On the flip side, mentors are instructed to recognize that postdoctoral fellowships are a training period — not an opportunity to obtain and exploit inexpensive labour. Mentors, too, should establish timelines for research as well as career development goals with their fellows, the compact says. They should base their relationship with fellows on mutual trust and respect. And they should ensure that postdocs have the opportunity to obtain the skills they need — whether on or off the bench.

These guidelines, although challenging, seem to be mutually beneficial. But the compact ups the ante for mentors by asking them to help postdocs explore career options outside academia and to commit to helping their mentees succeed on whatever path they choose. If both postdocs and mentors sign up to this compact it will require a lot of work and commitment from both sides. But it could also help end many of the complaints that each side has about postdoctoral training arrangements.

Paul Smaglik, *Naturejobs* editor

Editor: Paul Smaglik
Assistant Editor: Gene Russo

European Head Office, London
The Macmillan Building,
4 Crinan Street,
London N1 9XW, UK
Tel: +44 (0) 20 7843 4961
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

European Sales Manager:
Andy Douglas (4975)
e-mail: a.douglas@nature.com
Business Development Manager:
Amelie Pequignot (4974)
e-mail: a.pequignot@nature.com
***Nature*events:**
Claudia Paulsen Young (+44 (0) 20 7014 4015)
e-mail: c.paulsenyoung@nature.com
France/Switzerland/Belgium:
Muriel Lestringuez (4994)

UK/Ireland/Italy/RoW:
Loredana Milanese (4944)
Nils Moeller (4953)
Scandinavia/Spain/Portugal:
Evelina Rubio-Morgan (4973)
Germany/Austria/The Netherlands:
Reya Silao (4970)
Online Job Postings:
Matthew Ward (+44 (0) 20 7014 4059)

European Satellite Office
Germany: Patrick Phelan
Tel: +49 89 54 90 57 11
Fax: +49 89 54 90 57 20
e-mail: p.phelan@nature.com

Advertising Production Manager:
Stephen Russell
To send materials use London address above.
Tel: +44 (0) 20 7843 4816
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

***Naturejobs* web development:** Tom Hancock
***Naturejobs* online production:**
Catherine Alexander

US Head Office, New York
75 Varick Street, 9th Floor,
New York, NY 10013-1917
Tel: +1 800 989 7718
Fax: +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

Japan Head Office, Tokyo
Chiyoda Building, 2-37 Ichigayatamachi,
Shinjuku-ku, Tokyo 162-0843
Tel: +81 3 3267 8751
Fax: +81 3 3267 8746

Asia-Pacific Sales Manager:
Ayako Watanabe
e-mail: a.watanabe@natureasia.com

Lost in translation

English is the language of science. So to what extent are researchers who are non-native English speakers at a disadvantage? **Bonnie Lee La Madeleine** talks to scientists hailing from Japan to Germany.

The nervous Japanese postdoc spent two weeks creating slides, 30 hours drafting a script and 44 hours rehearsing. Altogether, she spent one month away from the bench so that she would not disappoint her supervisors and colleagues during a short informal presentation, in English, before co-workers. Yet they remembered only the mistakes, she says.

Seasoned scientists also feel under pressure when speaking in English. Masahiko Takada at the Tokyo Metropolitan Institute for Neuroscience admits that, even after years of working in English, "I sometimes feel frustrated when I have to discuss research data with foreign scientists."

Language mastery, be it of one's native or adopted tongue, provides the communicative ease that says: "I am capable." In science, weak English hinders a successful career. Improve your English proficiency, and confidence will follow — or so the people of many non-English-speaking nations believe.

Concerns about the dominance of the English language in science are being raised around the world. Researchers in Germany and France, for example, are grumbling about the frustration of working and publishing in English — and, perhaps more surprisingly, so are those in nations that have typically been viewed as consumers of basic science, rather than contributors.

A recent study in South Korea estimated how much an English-dominated setting for science has cost that nation's scientific development. Kumju Hwang of the University of Leeds, UK, interviewed 15 Korean researchers and engineers working in the United Kingdom about their personal experiences in the international arena. All the respondents said that because of language issues they spend a large proportion of their time preparing presentations and papers, and practising language skills for discussion — and even then, they say, they still miss more than 50% of what they hear. Seven of the interviewees felt that this weakness contributed to Korea's status as a consumer of basic science, rather than a major contributor. Hwang also found evidence that the requirement for English in science communication shaped the way that the social hierarchy among scientists developed in Korea. Employers prefer to hire researchers who have studied or worked in English-speaking nations, rather than those who did their postdoctoral work in countries such as Japan or France.

"Articles published in English only may be good for Japanese scientists proficient in English, but are not necessarily good for Japan."
— Masahiko Takada

Being misunderstood

Japan, on the other hand, has a history of translating science into and out of Japanese, so linguistic hurdles are accepted. Most Japanese researchers, when asked about such handicaps, will typically say that the language of science is English. Although nearly 75% of the 400 life-science journals published domestically in Japan are written in Japanese, these target applied rather than basic science researchers.

Basic science in Japan is becoming increasingly 'English only', as Japanese-language publications dedicated to those subjects disappear. RIKEN, one of Japan's most comprehensive groups of research facilities, has announced that its scientists published just under 2,000 original reports in English in 2005, and only 174 in Japanese. In Japan, a nation where English is the current language of knowledge production, domestic science society meetings are also moving towards English.

Internationally, reports expressing public concerns about adopting a single language for science appear periodically from different fields and nations — including those with far greater prowess in English than the Japanese, such as Spain, Germany and Portugal. Such essays and editorials lament the possible negative consequences of English-language dominance on national unity and economic stability in science and business.

"Insecurity in English is a widespread phenomenon," said Ulrich Ammon, professor of German linguistics at the University of Duisburg, echoing Takada's sentiment. "No one German is

M. KIMURA

entirely comfortable speaking and writing in English."

Ammon, who edited a 2001 compilation of essays called 'The Dominance of English as the Language of Science: Effects on Other Languages and Language Communities', adds that lingering resentments exist in "those nations that once held strong positions and were home to international languages of science". In France and Germany, the older scientists still remember the days when they could command an international readership in their native language. Ammon suggests that the shock of being dethroned, combined with the growing frustrations of misunderstanding and being misunderstood, has frustrated some researchers.

Like Hwang's report, these essays argue that non-native English speakers working in science are disadvantaged. "Nationally, everything moves more slowly," says Ammon. "It takes longer to write, communicate and respond, and there is a higher risk of misunderstanding." The problem is exacerbated by a greater likelihood of being ignored if your English is poor, he says.

Social inequality

Some European scholars have spoken out against the switch to English. A 2003 Finnish editorial warned that adopting English in Finland would alienate the lay people from the products of science. In this rallying cry, three Finnish academics contend that if university research focuses exclusively on the use of English, their own language will "gradually lose its ability to depict new concepts and phenomena and their subtle differences". They fear that this trend could create social inequality between those who can and cannot speak English.

In Spain, library science researchers Maria Bordons and Isabel Gomez at the Centre for Scientific Information and Documentation in Madrid used data from a survey of Spanish publication trends to try to understand how incentives designed by the Spanish government to aggressively increase the use of English in science are affecting researchers' publication preferences.

Bordons and Gomez found that reports on basic science, particularly in molecular biology and immunology, were published predominantly in English journals, whereas those on applied science were published in national journals in Spanish — mirroring the situation in Japan. They argued that these non-English journals, which are on the decline for both basic and applied science, are vital for knowledge transfer at the national level.

Two groups within the World Health Organization (WHO) issued a joint statement in 2004 asserting that non-English-speaking scientists prefer to be published in international journals. According to a study on publication preferences in addiction research, published by a different WHO body, this may be problematic in research fields where local traditions and experiences factor into a study.

According to Eugene Garfield, the inventor of the Science Citation Index, which tracks journal articles around the world via citation counts, local journals written in the native tongue can still play a part in an international setting. But "the burden of that communication belongs to the researchers," Garfield says. "It is up to the individual to decide when and how often it is necessary to translate his or her findings for local consumption."

Minoru Kimura (top) believes Japanese scientists must work at communicating in English; Ulrich Ammon argues that non-native English speakers will almost always be disadvantaged in science.

In Japan, review publications written by Japanese experts for non-English-speaking researchers and engineers have driven technology transfer since the seventeenth century. Hitoshi Okamoto, a lab head at RIKEN, says that these national-language publications exist to cement ideas and lubricate discussions between scientists and technicians. If such journals are indeed the locus of local knowledge transfer, then the disappearance of national-language publications in Spain is a concern.

Move to English

"Japanese scientists must work to communicate and write in English," says Minoru Kimura, president of the 2006 Annual Meeting of the Japan Neuroscience Society, and a researcher at Kyoto Prefectural University of Medicine. "But they must also present findings in a logical and attractive fashion, which seems independent of language." Perhaps, as Garfield suggests, it is a cultural, not linguistic, restraint that confines the communicative reach of Japan's researchers.

To build that communicative confidence, Japanese is slowly being eliminated from Japan's primary scientific content. Discussions in many of its newly established research institutes, and some university laboratories, are supposed to be conducted in English. According to white-paper reports of several Japanese institutes, as well as reports from Japanese ministries that oversee science and technology, this move to English aims to attract increased international attention and participation. For faster-paced interaction in competitive international settings, the increased exposure to English is beneficial — especially for younger researchers.

For this reason, Japan's more ambitious science societies are also moving to English-only. It is a move that is reluctantly accepted. "There are members who do not support this change," says Kimura. "They argue that presentations and subsequent discussions in English at the Japan Neuroscience Society annual meeting are less active than those in Japanese."

Society journals are also switching to English to make Japanese research accessible to scientists from other countries. This development, too, has its critics. Articles published in English only "may be good for Japanese scientists who are proficient in English, but this is not necessarily good for Japan", says Takada.

Meanwhile, Japanese scientists must budget for translation costs and subsidized language and communication training to increase exposure to, and proficiency in, English. Editing companies charge researchers US$500 to $800 per manuscript. Language training can cost $2,000 for a ten-week course, or about $50 per hour for a private lesson. These costs are additional burdens, and slow down scientific activity in the laboratory. Yet some science facilities, most notably the three RIKEN institutes dedicated to life sciences, are adopting English in all scientific activities, including administration.

"It is important to use English as an official language, especially for international participants," says Takada. "However, it is critical to use Japanese for effective communication in, for example, a committee, to ensure that important decisions are sufficiently considered." ■

Bonnie Lee La Madeleine is programme coordinator at RIKEN Brain Science Institute in Wakoshi, Japan.

The inside track from academia and industry

A degree of professionalism

There's a growing career path for students who like science, but don't want to be academics.

Michael S. Teitelbaum & Virginia T. Cox

Concerns about the adequacy of a country's scientific workforce feature in nearly all public discussions about 'innovation' and 'competitiveness'. Recently, for example, a wave of reports from industry associations and scientific associations called for substantial increases in public funding to expand the pipeline that ends with the production of newly minted PhD scientists.

The disparities in the claims are truly astonishing — all the way from shortages to oversupply of graduate scientists; and from claims by employers that they cannot find the recruits they need, to claims by junior scientists that they face dispiriting employment opportunities. The data are weak, anecdotal and often based on problematic labour-market forecasts. Typically, they do not account for disciplinary differences in supply and demand (see *Nature* **445,** 121, 124; 2007).

There is also concern about the content of graduate education in science. Employers often comment that scientists who want to work outside academia need more than strong scientific backgrounds. Graduate students, they say, should also seek skills in marketing, business and communication.

A fraction of science PhDs may indeed find additional professional training useful, again varying greatly by discipline and sector of employment. But a large proportion of those in scientific occupations outside academia do not have PhDs, and recruiters in many industries say that a large fraction of their recruits do not need PhDs, and may even be better off without them.

It is for such career paths, in science but outside academe, that professional science masters (PSM) degrees have been explicitly designed (see www.sciencemasters.com). There are now more than 100 such degree programmes, developed by science faculty at more than 50 US universities (in the interests of full disclosure, we should add that the Sloan Foundation has provided start-up grants to many of these programmes). Similar degree programmes are emerging in other countries.

These degrees are by no means clones of one another, but typically they are two-year 'science-plus' graduate degrees focusing primarily on courses, with less attention to lengthy research projects. Most are designed and taught by graduate science departments, and the coursework required often includes the bulk of the courses required of PhD candidates.

The 'plus' elements focus on intensive coursework and real-world experience in the business skills mentioned above — and such coursework attracts interest from PhD candidates as well. In addition, PSM programmes typically have close relationships with local employers seeking recruits with skills in science, business and management.

"Recruiters in many industries say that a large fraction of their recruits do not need PhDs, and may even be better off without them."

So far, at least, PSM programmes have given graduates attractive new pathways to science and engineering careers outside academe. Testimonials from PSM graduates suggest that early cohorts have found very appealing employment opportunities, better than baccalaureate holders and often as good as PhDs. Still, it is too early to know for sure how successful PSM graduates ultimately will be — many PSM courses are new and the number of graduates is relatively small, so it will take several more years before it is clear how their careers have progressed after initial employment.

The PSM is not in competition with the PhD. On the contrary, it is designed for those who are good at, and enthusiastic about, science and mathematics, but who do not wish to pursue the PhD-plus-postdoc route required for a job in academic research. Put another way, many undergraduates and graduate students deterred by the prospects for traditional careers in science may find themselves drawn to the PSM pathway.

If so, the availability of such a graduate career pathway might increase retention of freshmen in science and mathematics. According to the US National Science Board's 2004 figures, about half of the students who begin university intending to major in these subjects shift to other disciplines.

Sceptics might argue that PSM degrees run counter to the interests of research universities and their research-oriented faculty members — that PhD students are needed as research assistants for grant-supported research, and as teaching assistants to enable faculty members to devote more time to their externally funded research. For some institutions and some faculty members, such concerns may be justified.

Yet many institutions and science faculty members see one of their responsibilities as providing high-quality graduate science education for the non-academic workforce. This kind of perspective seems especially strong at leading public research universities, and at 'masters-focused' and comprehensive universities with strong science traditions. ■

Michael S. Teitelbaum is vice-president and Virginia T. Cox is programme associate, Alfred P. Sloan Foundation, 630 Fifth Avenue, New York, New York 10111. Comments welcome to teitelbaum@sloan.org